Unsupervised organization of image collections: taxonomies and beyond

Evgeniy Bart, Max Welling, and Pietro Perona

Abstract—We introduce a non-parametric Bayesian model, called TAX, which can organize image collections into a tree-shaped taxonomy without supervision. The model is inspired by the Nested Chinese Restaurant Process (NCRP) and associates each image with a path through the taxonomy. Similar images share initial segments of their paths and thus share some aspects of their representation. Each internal node in the taxonomy represents information that is common to multiple images. We explore the properties of the taxonomy through experiments on a large ($\sim 10^4$) image collection with a number of users trying to locate quickly a given image. We find that the main benefits are easier navigation through image collections and reduced description length.

A natural question is whether a taxonomy is the optimal form of organization for natural images. Our experiments indicate that although taxonomies can organize images in a useful manner, more elaborate structures may be even better suited for this task.

Index Terms—Taxonomy, hierarchy, clustering.

1 Introduction

Biological systems must deal with large numbers of images and image categories. Humans, for example, are familiar with tens of thousands of visual categories [1]; the number of individual objects and images is probably even higher. Computers are now starting to face the same challenge, due to the decreasing cost of acquiring, storing and distributing images.

Dealing efficiently with so many images requires organizing them into an appropriate structure. Organizing images is useful for several reasons. Organization provides a compact and efficient representation of image collections and facilitates navigation of these collections, which speeds up searching and thus learning and recognition. Additional benefits, although not explored in this paper, may include performing recognition at different levels of specificity, and forming appropriate priors for learning new categories [2], [3].

It is unclear what is the best form of organization. Applications such as navigating and searching through a collection of images suggest using tree-shaped hierarchies, or taxonomies (Figure 1). The reason is that taxonomies are known to reduce the complexity of search. Taxonomies may underlie the 'natural' statistics of many things that surround us. Philosophers interested in ontology recognized early on that the world may be organized into mearningful hierarchies. Taxonomies were used successfully to organize very large numbers of living species (e. g. the work of Linnaeus) as well as the contents of the web (Yahoo! Directory). Statisticians have recently suggested methods for assigning priors to taxonomies and for inferring taxonomies from data [4]. This study is an exploration of the power of these methods in the context of visual taxonomies. In our experiments we do indeed find that the taxonomies we obtain from TAX do speed up browsing of an image collection considerably. The question of whether taxonomies are the best organization of an image collection brings us to consider possible organizations beyond taxonomies in section 4.3.

The organizations explored in this paper are based only on visual properties. Although organizations based on image semantics could be useful as well, we show that even purely visual taxonomies can facilitate several visual tasks.

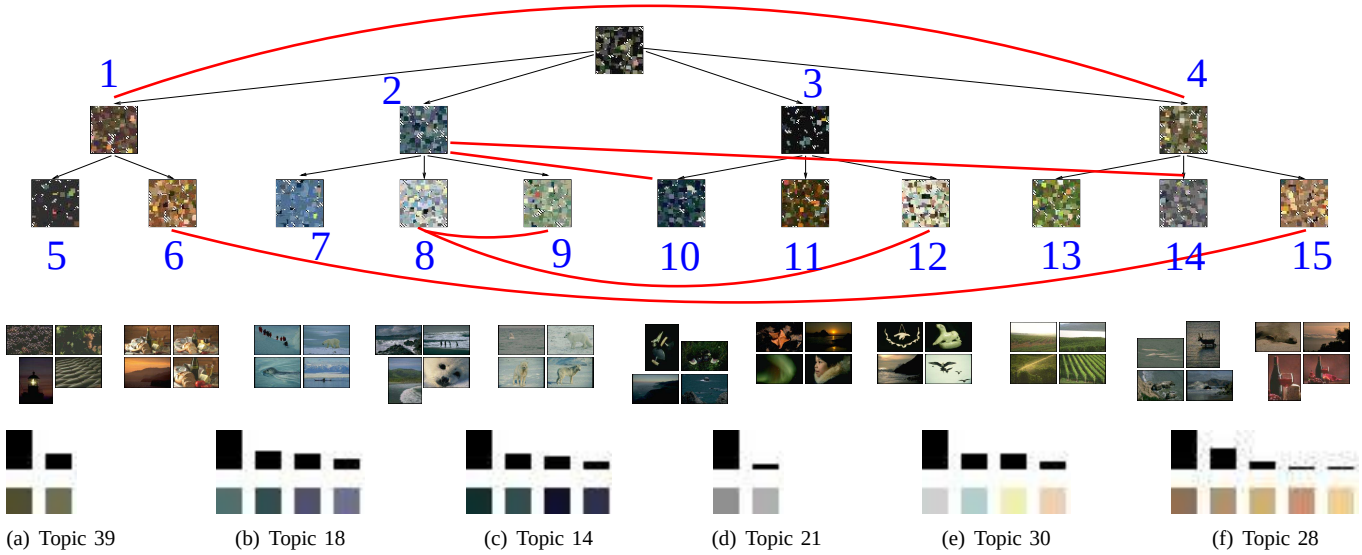
The remainder of this paper is organized as follows. In section 2 we briefly review the relevant literature. In section 3 we present the proposed TAX model. Experiments with this model are described in section 4. We discuss the conclusions from these experiments in section 5.

- *E. Bart and P. Perona are with California Institute of Technology, Pasadena, CA, 91125.*
  E-mail: {bart, perona}@caltech.edu
- *M. Welling is with UC Irvine, Irvine, CA, 92697.*
  E-mail: welling@ics.uci.edu

2 Brief survey of previous work

There are several lines of research on organizing image categories into taxonomies in a supervised manner [5], [6], [7]. In [8], a manually constructed hierarchy of labels of image segments was exploited to improve image labeling. This research indicates that organizing categories is useful. It is therefore of great interest to attempt an unsupervised organization of images as well, because uncategorized images are abundant in everyday life. An additional benefit from unsupervised organization is that it might hint at how natural categories arise in the first place.

Clustering [9] has been the dominant method for unsupervised image organization. In clustering, images are organized into a flat list of groups. The number of groups may be fixed in advance, or may be selected automatically [4], [10]. A potential disadvantage of clustering is that it may not be flexible enough for some datasets. For example, no single number of clusters may be appropriate for a given image collection and



(a) Topic 39 (b) Topic 18 (c) Topic 14 (d) Topic 21 (e) Topic 30 (f) Topic 28

Fig. 1. Primary taxonomy structure and secondary graph structure on a dataset of 300 Corel images (section 4.1.1). This taxonomy is purely based on the color of image patches, disregarding position and other information. Top: the taxonomy. Each node shows a synthetically generated 'quilt' – an icon that represents that node's model of images (section 4.1.1.1). As can be seen, common colors (such as black, green or blue) are represented at top nodes and therefore are shared among multiple images. The black edges represent the original taxonomy structure. The thick red edges connect nodes that are not in a hierarchical relationship with each other, but nevertheless share a topic. Below each leaf, four most probable images from that leaf are shown. Bottom: the shared topics. Topic 39 is shared between nodes 1 and 4; topic 18 is shared between nodes 2 and 14; topic 14 is shared between nodes 2 and 10; topic 21 is shared between nodes 8 and 9; topic 30 is shared between nodes 8 and 12; and topic 28 is shared between nodes 6 and 15. As can be seen, topics are shared between categories that have common aspects in their appearance. Moreover, it can be seen that the shared topics indeed represent the shared aspects of appearance.

a given task. Using few clusters may force dissimilar images together and thus discard important distinctions between the images. Conversely, using more clusters may be inefficient (in terms of speed and memory). In contrast, a taxonomy allows to simultaneously exploit the advantages of few clusters at the top levels of the hierarchy and of many clusters at the bottom levels.

In addition, flat clustering does not explicitly represent similarities between the clusters. (One exception is the TDP model [11], where the clusters are different transformed versions of a smaller number of meta-clusters.) But similarities between some natural categories (such as cats and dogs) do exist, and have been successfully exploited in the past [3], [12], [2], [13], [14]. This suggests that an organization that preserves some information about distance may be beneficial, and it's tempting to obtain such an organization in an unsupervised manner.

Several methods of organizing documents into taxonomies in an unsupervised manner have been developed in the past. Examples include [15], where a manually specified taxonomy (a binary tree) was used for modeling visual and text features jointly, and several hierarchical clustering schemes (e. g. [16]). More recently, a model capable of learning the taxonomy structure was proposed. The model is called 'Nested Chinese Restaurant Process' (NCRP), and it was introduced in [17], with applications to text data. NCRP and similar models were successfully applied to visual data in [18], [19]. Our model, called TAX, is similar to NCRP, and is based on [18]. The key differences with NCRP are discussed in section 3.1.

3 Learning Visual Taxonomies

In our model, as in NCRP, the taxonomy is defined by how information is shared among multiple images. The main idea is that information common to multiple images is represented only once at a node common to these images. The technical details of this process are described below.

3.1 TAX: The generative model

We approach taxonomy learning in the framework of generative modeling. In our model, images are generated by descending down the branches of a tree. The shape of the tree is estimated from a collection of training images. Thus, model fitting produces a taxonomy for a given collection of images. The model (called TAX) is summarized in Figure 2.

Images in TAX are represented as unordered collections of local features (this is known as the 'bag-of-words' representation). The basic blocks of this representation are called 'detections'. Each detection represents the visual appearance of a small image region. Typically, these regions are centered at points detected by an interest operator, hence the term 'detection', although regions centered on fixed grids have been shown to have equivalent descriptive power [20]. We denote by D_i the total number of detections in image i. The numbers D_i are observed (i. e., known in advance) and may be different for different images.

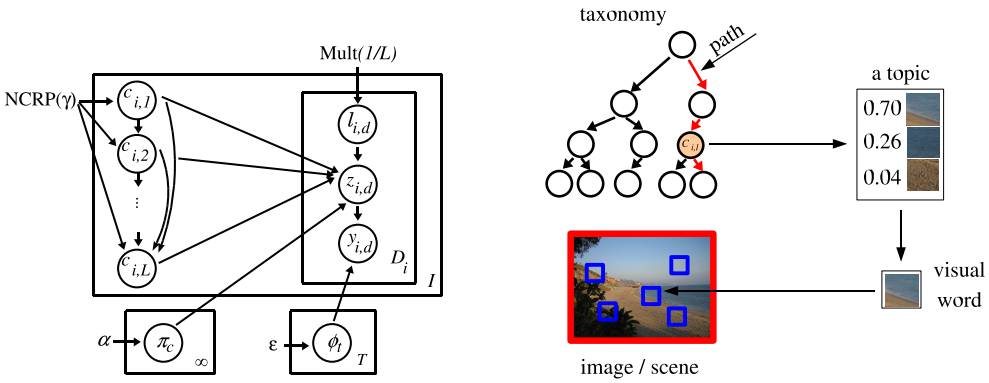


Fig. 2. Left: TAX, the generative model for learning visual taxonomies. Right: an illustration of the generative process. An image i is generated as follows. First, a complete path from the root to a leaf through the current taxonomy is sampled. A sample path is highlighted by the red arrows in the figure. Since the hierarchy depth is fixed, the path has length L. The ℓ'th node on this path is denoted $c_{i,\ell}$ (see plate diagram on the left). For every detection d in the image i, we sample $\ell_{i,d}$ – a level in the taxonomy – from a uniform multinomial distribution over the L nodes on this path. The node from which this detection is generated is then $c = c_{i,\ell_{i,d}}$. We then pick a topic $z_{i,d}$ from π_c – the distribution over topics at that node. Finally, we pick a detection $y_{i,d}$ from the multinomial distribution $\phi_{z_{i,d}}$, associated with topic $z_{i,d}$.

The description of detections' visual appearance is quantized into a number of distinct classes, and each class is called a 'visual word'. The visual dictionary is the set of all visual words used by the model, and its size is denoted by W. This dictionary can be pre-specified or learned from training data.

Similarly to LDA [21], [22], [20], distinctive patterns of co-occurence of visual words are represented by 'topics'. A topic is a multinomial distribution over the visual dictionary. In many cases, sparse distributions are desirable, so that only a subset of visual words have substantial probability in a given topic. Thus, a topic represents a set of words that tend to co-occur in images. Typically, this corresponds to a coherent visual structure, such as skies or sand [20]. The total number of topics in the model is fixed and denoted by T. Each topic ϕ_t has a uniform Dirichlet prior with parameter ϵ.

A category is represented as a multinomial distribution over the T topics. The distribution for category c is denoted by π_c and has a uniform Dirichlet prior with parameter α. For example, a 'beach scenes' category might assign high probabilities to the 'sea', 'sand', and 'skies' topics.

In TAX, categories are organized hierarchically, as in Figure 1. For simplicity, we assume that the hierarchy has a fixed depth L (this assumption can be relaxed). Each node c in the hierarchy represents a category, and is therefore associated with a distribution over topics π_c. (This is the key difference from NCRP: in NCRP, each node in the taxonomy contains just one topic, rather than a distribution over topics.) The number of categories is thus not fixed a priori; rather, it is determined by the hierarchy structure (there are as many categories as there are nodes in the hierarchy).

Next, we describe how the taxonomy structure is generated. A nonparametric prior over trees of depth L, known as the 'nested Chinese restaurant process' (NCRP), is used [17]. This prior is flexible enough to allow learning an arbitrary taxonomy, but also allows for efficient inference. Briefly, the tree is determined implicitly by sampling a path for each image. The paths are sampled one-by-one. All paths start at the root of the tree. For the first image, the only option is a linear path of length L. For subsequent images, a sequence of $L-1$ random steps is made. At each step, the path can either follow an existing branch, or create a new branch. (Note that conceptually the tree has infinitely many branches at each node. By existing branches, we mean branches already traversed by at least one previous path. By new branches, we mean branches not yet traversed by any previous path.) The probability of following an existing branch is proportional to the number of previous paths that traversed that branch, and the probability of creating a new branch is proportional to the parameter γ:

$$p(\text{existing branch } c) = \frac{m_c}{\gamma + \sum_c m_c} \quad (1)$$

$$p(\text{new branch}) = \frac{\gamma}{\gamma + \sum_c m_c}. \quad (2)$$

Here m_c is the number of previous paths that took branch c.

An important property of the NCRP prior is that the probability of any image's path does not depend on the order in which the images were considered. This property is called 'exchangeability'. See [17] for details.

The TAX model could be used generatively to produce an image collection; this is not the inference process that will be used to produce a taxonomy, but describing the generative process will clarify the nature of TAX. The process is as follows. First, the paths for all images are sampled from the NCRP prior. Since the hierarchy depth is fixed, each path has length L. The ℓ'th node on the path for image i is denoted $c_{i,\ell}$, and it is sampled according to eqs. (1), (2). These paths implicitly determine the tree. The topics and categories are sampled from Dirichlet priors: $\pi_c \sim \text{Dir}^T(\alpha)$, $\phi_t \sim \text{Dir}^W(\varepsilon)$.

Next, the detections for all images in the collection are generated. To sample detection d in image i, we first sample

$\ell_{i,d}$—a level in the taxonomy—from a uniform multinomial distribution over the nodes on the image's path: $\ell_{i,d} \sim \mathrm{Mult}(1/L)$. This determines a node on the path for image i—namely, the node $c = c_{i,\ell_{i,d}}$. We then pick a topic $z_{i,d}$ from π_c—the distribution over topics at that node—according to $z_{i,d} \sim \mathrm{Mult}(\pi_{c_{i,\ell_{i,d}}})$. Finally, we pick a detection from the multinomial distribution $\phi_{z_{i,d}}$ associated with topic $z_{i,d}$: $y_{i,d} \sim \mathrm{Mult}(\phi_{z_{i,d}})$. The resulting joint probability is thus

$$\begin{aligned} p \;=\; & \prod_{i=1}^{I}\prod_{\ell=1}^{L} p_{\text{NCRP}}(c_{i,\ell}|\gamma, \{c_{p,q}\}_{(p<i)\vee(p=i\wedge q<\ell)}) \cdot \\ & \cdot \prod_{t=1}^{T} \mathrm{Dir}^W(\phi_t|\varepsilon) \cdot \prod_{c=1}^{\infty} \mathrm{Dir}^T(\pi_c|\alpha) \cdot \\ & \cdot \prod_{i=1}^{I}\prod_{d=1}^{D_i} \mathrm{Mult}(\ell_{i,d}|1/L)\,\mathrm{Mult}(z_{i,d}|\pi_{c_{i,\ell_{i,d}}}) \cdot \\ & \cdot \mathrm{Mult}(y_{i,d}|\phi_{z_{i,d}}), \end{aligned} \tag{3}$$

where I is the total number of images, and $p_{\text{NCRP}}(c_{i,\ell}|\gamma, \{c_{p,q}\}_{(p<i)\vee(p=i\wedge q<\ell)})$ is the probability of sampling the node c given all previously sampled nodes (that is, all nodes for previous images, and the previous nodes for the current image). Figure 2 illustrates the generative process.

Note that in TAX (as in NCRP), the probability of an image with more detections (larger D_i) involves more terms (such as π and ϕ) in the last product in eq. (3). Since these terms are probabilities and thus smaller than 1, the total probability of an image with more detections is smaller than the probability of an image whose detections are similarly distributed but fewer in number. Images with fewer detections are thus intrinsically more probable. This could be problematic, in particular, for the navigation application, and the implications are described in section 4.2.1. But for many applications, such as inferring the taxonomy itself, as well as categorization, this is not a problem. The reason is that in these applications we do not compare the probabilities of two different images. Instead, we compare the probability of the same image under different models (such as different paths through the taxonomy or different category models). In our experiments, to simplify this aspect, the image has a fixed number of detections; thus, the number of terms is also fixed and does not influence the comparison.

Recall that our criterion for a useful taxonomy is that shared information is represented at nodes which are higher up in the tree and are shared among many images. The generative process described above is naturally suited to this criterion. The nodes higher up in the taxonomy are used by many paths; the information they represent is therefore shared by many images. For instance, the root node is necessarily used by all paths and therefore will model very general topics that exist in all images. Conversely, the lower a node is in the taxonomy, the fewer images traverse it, and the more image-specific the information at that node is.

Compared to the original NCRP model [17], the proposed TAX model allows representing several topics at each node in the taxonomy. In addition, it makes all topics available at every node. Although NCRP has been used successfully in text modeling [17], we found experimentally that these changes were necessary to infer visual taxonomies. This is consistent with [19], who also used a modified version of NCRP for image data.

3.2 Inference

The goal of inference is to learn the structure of the taxonomy and to estimate the parameters of the model (such as π_c) given an image collection and the parameters α and ε of the prior densities. The overall approach is to use Gibbs sampling, which allows drawing samples from the posterior distribution of the model's parameters given the data. Taxonomy structure and other parameters of interest can then be estimated from these samples. Compared to the sampling scheme used for NCRP [17], we augmented Gibbs sampling with several additional steps to improve convergence. The details are given below, but the remainder of this section may be skipped on first reading.

To speed up inference, we marginalize out the variables π_c and ϕ_t. Gibbs sampling then produces a collapsed posterior distribution over the variables $\ell_{i,d}$, $c_{i,.}$, and $z_{i,d}$.

The paths $c_{i,.}$ (and thus the taxonomy structure) are initialized at random from the NCRP prior (eqs. (1) and (2)). The remaining variables ($\ell_{i,d}$ and $z_{i,d}$) are initialized uniformly at random. Several additional initializations were tried, including assigning all images to the same linear path and assigning each image to an individual unique path. We have also tried initializing the tree structure to form a binary tree. This binary tree was created by hierarchical multinomial k-means with $k = 2$. At each node. we split all images in that node into two groups by multinomial k-means, and created two child nodes populated by images from each group. This procedure was repeated recursively until the desired number of levels was created. The topics were initialized to the multinomial cluster centroids. The results with all these initializations were similar.

The initialization described above produces a taxonomy, with all images assigned to some paths, and all detections assigned to some level and some topic in that level. Of course, this taxonomy is random. The next step is to sample the individual variables until the taxonomy converges to a good local optimum. This sampling is performed in iterations, or sweeps, with each iteration corresponding to resampling all variables in the model once. The variables include the tree paths $c_{i,.}$ and the level/topic variables $\ell_{i,d}$ and $z_{i,d}$.

To resample a path for an image, this image is first removed from the taxonomy. If a particular node was used only by this image, the node is removed as well. Generally, most nodes in the taxonomy are left intact since a single image only uses L nodes, and many of these are shared. Note also that removing an image may change the tree structure, but the taxonomy still remains a tree. This is because if a node is removed, all of its subnodes are guaranteed to be removed as well.

Next, we consider all possible paths for this image. These may include existing paths as well as new paths. We calculate the probability of the image in all of these paths (eq. (6)) and sample a path from this distribution.

To resample $\ell_{i,d}$, the detection is first removed from the tree. Then the L possible new assignments are considered, and the

probability of each is calculated (eq. (4)). The new value for $\ell_{i,d}$ is sampled from this distribution. The $z_{i,d}$ variables are resampled similarly.

The following conditional distributions need to be calculated to perform this sampling: $p(\ell_{i,d} = \ell|\text{rest})$ (the probability of sampling a level ℓ for detection d in image i given values of all other variables), $p(z_{i,d} = z|\text{rest})$ (the probability of sampling a topic z for detection d in image i) and $p(c_{i,\cdot}|\text{rest})$ (the probability of sampling a path for image i through the current taxonomy; note that this includes the possibility to follow an existing path, as well as to create a new path). These conditional distributions, as usual, are expressed in terms of count values. The necessary counts are described next. $N_{i,\ell}$ is the number of detections in image i assigned to level ℓ. $N_{i,l,t}$ is the number of detections in image i assigned to level ℓ and topic t. $N_{t,w}^{\neg(i,d)}$ is the number of detections belonging to visual word w and assigned to topic t across all images, excluding the current detection d in image i. A dot in place of an index indicates summation over that index, so $N_{t,\cdot}^{\neg(i,d)}$ is the total number of detections assigned to topic t (excluding the current detection d in image i). $m_c^{\neg i}$ is the number of images that go through the node c in the tree, excluding the current image i. $N_{c,t}^{\neg(i,d)}$ is the number of detections assigned to node c and topic t, excluding the current detection (i,d). $N_{c,t}^{\neg i}$ is the number of detections assigned to node c and topic t, excluding all detections in the current image i. Finally, $N_{c,\cdot}^{\neg(i,d)}$ is the total number of detections assigned to node c, excluding the current detection (i,d), and $N_{c,\cdot}^{\neg i}$ is the total number of detections assigned to node c excluding all detections in image i. In terms of these counts we can derive the following conditional distributions:

$$p(\ell_{i,d} = \ell|\text{rest}) \propto \frac{\alpha + N_{c_{i,\ell},z_{i,d}}^{\neg(i,d)}}{\alpha T + N_{c_{i,\ell},\cdot}^{\neg(i,d)}} \tag{4}$$

$$p(z_{i,d} = z|\text{rest}) \propto \left(\alpha + N_{c_{i,\ell_{i,d}},z}^{\neg(i,d)}\right) \cdot \frac{\epsilon + N_{z,y_{i,d}}^{\neg(i,d)}}{\epsilon W + N_{z,\cdot}^{\neg(i,d)}} \tag{5}$$

$$p(c_{i,\cdot} = c|\text{rest}) \propto \prod_{\ell} \left(m_{c_{i,\ell}}^{\neg i}\mathbb{I}[m_{c_{i,\ell}}^{\neg i} > 0] + \gamma\mathbb{I}[m_{c_{i,\ell}}^{\neg i} = 0]\right) \times \prod_{\ell} \frac{\prod_t \Gamma(\alpha + N_{c_{i,l},t}^{\neg i} + N_{i,\ell,t})}{\Gamma(\alpha T + N_{c_{i,\ell},\cdot}^{\neg i} + N_{i,l})} \frac{\Gamma(\alpha T + N_{c_{i,\ell},\cdot}^{\neg i})}{\prod_t \Gamma(\alpha + N_{c_{i,\ell},t}^{\neg i})} \tag{6}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

The first two equations have intuitive meaning. For example, eq. (5) consists of two terms. The first term is (up to a constant α) proportional to $N_{c_{i,\ell_{i,d}},z}^{\neg(i,d)}$. Here $c = c_{i,\ell_{i,d}}$ is simply the category node to which the detection in question (namely, detection d in image i) is currently assigned. Thus, $N_{c,z}^{\neg(i,d)}$ is just the number of other detections already assigned to topic z. The topics thus have a clustering property: the more detections are already in a topic, the more likely another detection is to be assigned to the same topic. The second term in eq. (5) is (again, up to the prior ϵ) the fraction of instances of the current visual word (e. g. visual word 3 if $y_{i,d} = 3$) among all visual words in the topic z. This term encourages detections which are highly probable under topic z, and penalizes those which are improbable. Overall, eq. (5) is quite similar to the corresponding equation in standard LDA [21].

The last equation (eq. (6)) is harder to understand, but it's quite similar to the corresponding equation in the NCRP [17]. The first term represents the prior probability of generating the path c given all other paths for all other images according to the NCRP prior. Note again that this prior is exchangeable: changing the order in which the paths were created does not change the total probability [17]. Therefore, we can assume that the current path is the last to be generated, which makes computing the first term efficient. The second term represents how likely the detections in image i are under the path c.

Additional details about inference are given in the Supplemental Material.

The collapsed sampler only estimates the ℓ, z, and c variables. Note that the c variables determine the tree, as described in section 3.1. Sometimes, it is also useful to have π_c and ϕ_t. These can be estimated as

$$\bar{\phi}_{t,w} = \frac{\epsilon + N_{t,w}}{\epsilon W + N_{t,\cdot}} \quad \text{and} \quad \bar{\pi}_{c,t} = \frac{\alpha + N_{c,t}}{\alpha T + N_{c,\cdot}}. \tag{7}$$

4 Experiments

Our experiments with unsupervised learning of visual taxonomies are divided into three groups. In section 4.1, some basic results are shown to illustrate the kind of taxonomies that are learned by our model. In section 4.2, we demonstrate that these taxonomies facilitate several visual tasks (namely, navigation and compression). Finally, in section 4.3 we show experiments suggesting that an organization more complex than a taxonomy might be necessary for natural images.

4.1 Basic experiments with taxonomies

4.1.1 Corel images and color features

Color is easily perceived by human observers, making color-based taxonomies easy to analyze and interpret. Our first experiment is therefore a toy experiment based on color. Experiments on more realistic datasets are reported below.

A subset of 300 color images from the Corel dataset (collections 14000, 144000, and 157000, selected arbitrarily) was used. This dataset will be called 'Corel 300' below. The images were rescaled to have 150 rows, preserving the aspect ratio. The visual words were pre-defined to represent images using 'color histograms'. Eight uniformly spaced bins for each of the three color channels were used. This resulted in a total of $8 \cdot 8 \cdot 8 = 512$ visual words, where each word represents a particular color (quantized into 512 bins).

500 pixels were sampled uniformly from each image and encoded using the color histograms. A TAX model with four levels ($L = 4$) and 40 topics ($T = 40$) was fitted to the data. The remaining parameters were set as follows: $\gamma = 0.01$, $\varepsilon = 0.01$, $\alpha = 1$. These values were chosen manually (one could explore ways of setting these values automatically [23]). Gibbs sampling was run for 1000 iterations, where an iteration corresponds to resampling all variables once. The number 1000 was selected by monitoring training set perplexity to determine convergence.

Fig. 3. Taxonomy learned on the dataset of 10,000 Corel images. The display is similar to Figure 1. Due to its size, the taxonomy was divided into five parts. The top node in each panel represents the same node – namely, the single root of the taxonomy. In other words, even though the root is repeated five times (once in each panel), there is only one root in the taxonomy. Each panel shows a subset of second-level nodes and all of their children. The remaining second-level nodes in each panel are shown at a reduced size, and their children are omitted.

The resulting taxonomy is shown in Figure 1. The main conclusions are that the images are combined into similar-looking groups at the leaves, and that the topics learned by the leaves represent colors that are dominant in the corresponding groups.

4.1.1.1 Computing quilt images: The quilt basically shows what the model 'thinks' images rooted at a given node look like. Recall that each node represents a distribution over topics, and each topic is itself a distribution over words. The quilt represents this pictorially, and is generated from the generative model learned at that node. We start with a blank image. A topic is sampled from the node-specific topic distribution. Then a word is sampled from that topic. This process is repeated 1000 times, to obtain 1000 word samples. Next, for every word a location in the 'quilt' image is sampled uniformly at random. That location is then painted with the color corresponding to the visual word (recall that each word represents a particular color). In practice, a 5×5 pixels patch is painted to make colors more visible.

4.1.1.2 Stability of the taxonomies: The TAX model (like most probabilistic models) admits multiple modes in the underlying probability distribution. Re-running Gibbs sampler multiple times may find these multiple modes. In our experiments, re-running Gibbs sampling 15 times yielded four significantly different taxonomies (corresponding to four different modes). These results will be discussed in more detail in section 4.3.1. The main conclusion here is that the inference procedure is stable and can reliably fit the model to the data.

4.1.1.3 Corel 10,000: The taxonomy learned in a similar manner from 10,000 images (a subset of 100 directories from the Corel collection, chosen at random and listed in the Supplemental Material) is shown in Figure 3. As can be seen, colors that are shared among multiple images appear higher up in the taxonomy. For example, black and white pixels are prevalent in all images and are represented at the root. Intermediate nodes (at the second level) represent colors shared by lower-level nodes. For example, in the top panel, nodes 1–6 contain red objects. The topic that represents red appears only once in the node A shared among all of nodes 1–6. Similarly, nodes 7, 8 and 10, and images 2 and 3 of node 9, share the shade of blue represented at their parent node B.

Note that many taxonomies are quite broad (i. e., each node has many children). This is controlled by the γ parameter of NCRP. The shape of the taxonomy is not very sensitive to γ: it has to be changed by a factor of 100 or more to have noticeable effect. We tried artificially restricting the maximum fan-out (the number of children per node) to four. This was achieved by modifying eq. (2) to assign zero probability to all branches beyond four. This resulted in somewhat poorer performance. For example, the navigation performance (section 4.2) on the Corel 300 dataset increased to 68, from 49 in the original model (lower values are better, see section 4.2.1). Increasing the number of levels in the restricted fan-out model to five brought the performance to 52. Increasing the number of levels beyond five did not improve performance further. The conclusion is that artificially restricting the fan-out of the model does not improve performance.

4.1.2 13 scenes dataset

In this section we describe an experiment carried out on a more challenging dataset of 13 scene categories [20]. We used 100 examples per category to train a taxonomy model. The size of the images was roughly 250×350 pixels. From each image we extracted 500 patches of size 20×20 by sampling their location uniformly at random. This resulted in 650,000 patches from which 100,000 were randomly selected. For these 100,000 patches SIFT descriptors [24] were computed and clustered by running k-means for 100 iterations with 1000 clusters. The centroids of these 1000 clusters defined the visual words of our visual vocabulary. Each of the 500 patches for each image (again, represented as a SIFT descriptor) was subsequently assigned to the closest visual word in the vocabulary. The proposed TAX model was then fitted to the data by running Gibbs sampling for 1000 iterations. We used four levels for the taxonomy and 40 topics. The remaining parameters were set as follows: $\gamma = 100$, $\varepsilon = 0.01$, $\alpha = 1$.

The taxonomy is shown in Figure 4. Roughly, the images are split into three groups: natural scenes (such as forest and mountains), cluttered man-made scenes (such as tall buildings and city scenes), and scenes with open spaces (such as highways or coasts).

4.1.3 Supervised taxonomy learning

If category labels are available, it is useful to take them into account when learning a model. Adding supervision to TAX is quite simple, and can be done in one of two ways. One possibility is, during Gibbs sampling, to simply disallow images of different categories to be in the same path. Another possibility is to modify the model in Figure 2 so that there is a single path $(c_1, \ldots, c_L)$ for every *category* of images, rather than for every image. In this case categories may share partial or complete paths. (This second possibility is equivalent to pooling together all detections from each category of images into a single bag of words.) We experimented with both methods. Their performance was very similar in terms of recognition rate, but the second method was more efficient in terms of running time, because fewer paths had to be resampled in every iteration of Gibbs sampling. The results below are therefore reported with this second method.

The taxonomy learned in the supervised manner is shown in Figure 5. As expected, most categories are separate. The only four categories that ended up in the same leaf node are office, bedroom, kitchen, and living room. This is consistent with previous results indicating that these categories are very difficult to distinguish [20], [25].

It is also of interest to look at categories that share intermediate nodes. From left to right, these categories are: suburb, mountain, and tall building; coast, highway, forest, inside city, and street; and open country, office, bedroom, kitchen, and living room. Note that these groupings are based on low-level visual properties such as texture similarity, rather than high-level semantic knowledge. Some of the groupings are similar to those in Figure 4. For example, coast and highway are grouped together in node B in Figure 4, and inside city and street are grouped together in node A in Figure 4. These groupings clearly are due to significant visual similarity

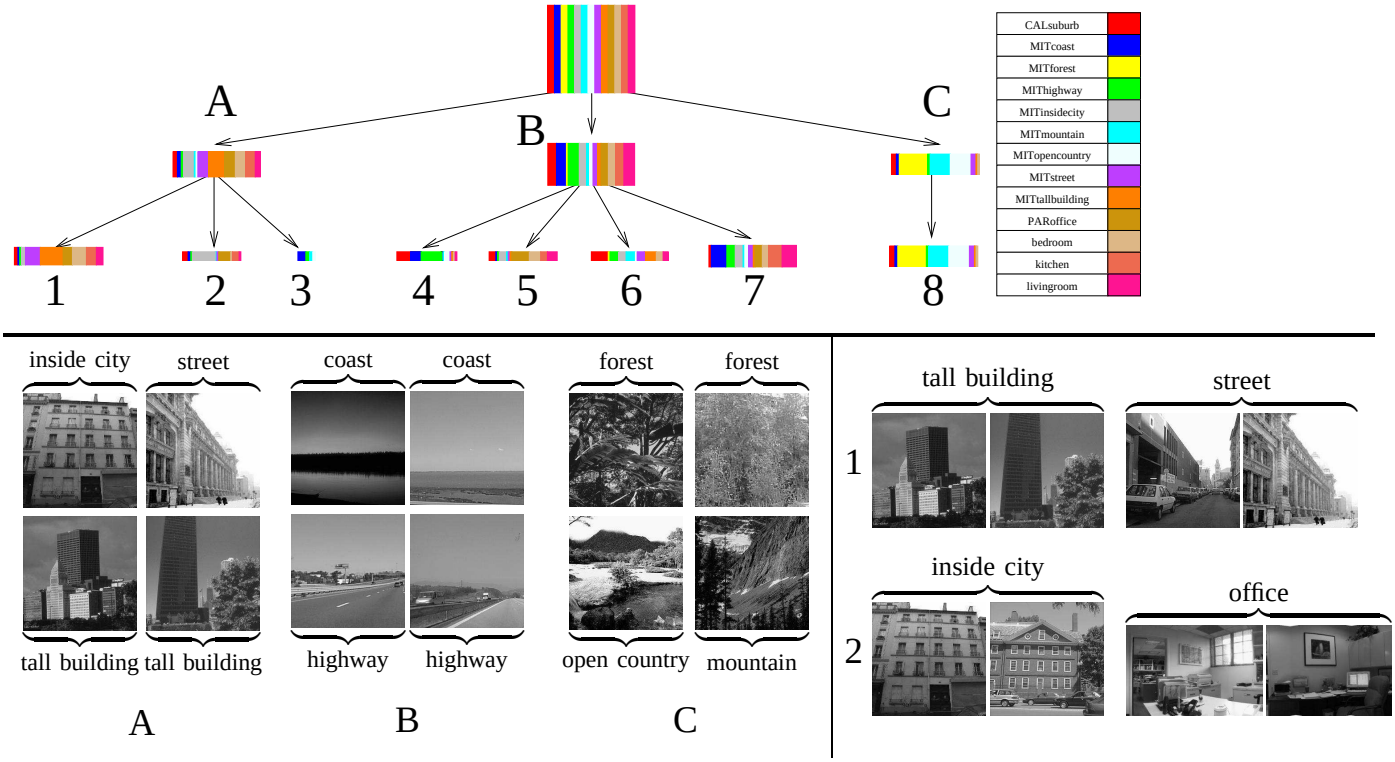


Fig. 4. Unsupervised taxonomy learned on the 13 scenes dataset. Top: the entire taxonomy shown as a tree. The size of each node is proportional to the number of images in that node. Each category is color-coded according to the legend on the right. The proportion of a given color in a node corresponds to the proportion of images of the corresponding category. Note that this category information was not used when inferring the taxonomy. There are three large groups, marked A, B, and C. Four images most probable under each group are shown. As can be seen, group A contains cluttered man-made scenes, such as tall buildings and streets. Group C contains cluttered natural scenes, such as forest and mountains. Group B contains uncluttered scenes, such as highways or coasts. These groups split into finer sub-groups at the third level. One such split is shown at the bottom right. For nodes 1 and 2, four images most probable under the corresponding node are shown. (Node 3 is not shown due to its small size.) As can be seen, group A splits into scenes containing extended sky areas (node 1) and scenes containing little or no sky (node 2). Each of the third-level groups splits into several tens of fourth-level subgroups, typically with 10 or less images in each. These are omitted from the figure for clarity.

between the categories in question. For example, the main difference between categories 'inside city' and 'street' is the viewing direction: along the street for 'street' and towards the buildings for 'inside city' (see top two pictures in Figure 4, node A). Of course, since supervised TAX uses additional information unavailable in unsupervised training, there are also some noticeable differences between the supervised and unsupervised versions of the taxonomies.

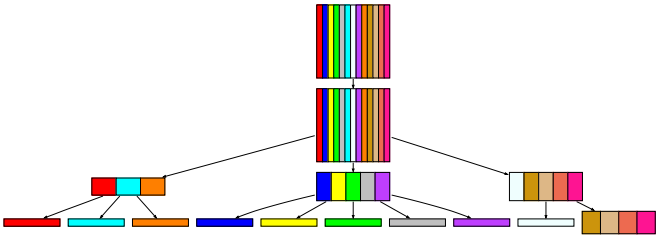


Fig. 5. Supervised taxonomy learned on the 13 scenes dataset. The taxonomy is shown as a tree, with each category color-coded according to the legend in Figure 4. The proportion of a given color in a node corresponds to the proportion of images of the corresponding category. Note that TAX decided to collapse all indoor scene categories into one node.

4.2 Usefulness of the taxonomies

In section 4.1, we have shown that the model learns taxonomies that are intuitively appealing. In this section, we show that the learned taxonomies are useful for representing images. We experiment with two tasks and show that both are facilitated by using taxonomies. The first task involves a user (human or computer) navigating through an image collection. The performance is measured by how quickly the user can find a specific image in a large collection. The second task is image compression. The performance is measured by the amount of memory necessary to represent a collection. These tasks will be described in more detail below.

To show that a hierarchical organization is advantageous

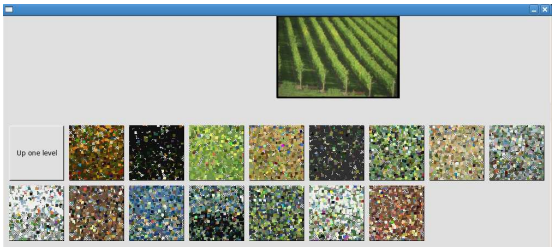


Fig. 6. Illustration of one step in the browsing process (section 4.2.1).

over a flat organization (such as standard clustering), we use the following strategy. We fit several taxonomies of varying depths to several datasets. We then perform each of the two tasks using each of the taxonomies and compare performance. In our experiments, relatively deep taxonomies (3–4 levels) generally achieve the best performance. In particular, they outperform the shallow two-level taxonomies (note that a two-level taxonomy is quite similar to a flat clustering).

4.2.1 Navigation using taxonomies

Hierarchical organization has long been used to deal with complexity. Therefore, it seems natural to use our taxonomies to help users and/or computers navigate through large image collections. By navigation, we mean finding an image of interest in the collection. In general, the user does not have this image of interest in advance. (If the image were available in advance, then methods for finding duplicates could be used instead.) A canonical example of this situation is finding a previously seen person, whose image is not currently available, in a database of mugshots.

A naive way to navigate the collection is to view all images one by one. This would require time linear in the number of images (on average, half of all the images in the collection will be considered until the target image is found). To make navigation more efficient, the images could be organized into groups. Navigation would then proceed by selecting an appropriate group first, and then selecting an image from that group. In the optimal situation (if the number of groups is equal to the number of images per group), this can improve the average navigation time to $O(\sqrt{n})$ (where n is the number of images). A well-organized taxonomy can be even more efficient and achieve $O(\log n)$ navigation performance.

For inspiration, we will consider below another task—selling stock photographs online. In this application, photographs matching a query (or, possibly, all photographs in a collection) are shown as a list, and the customer scrolls through them to find a relevant one. Again, note that the target image of interest is not available in advance. Typically, hundreds of photographs are present, and the search is time-consuming. To reduce the effort, the images can be organized into semantic groups, but such organization requires significant manual effort. We are interested in exploring whether the taxonomy learned automatically by TAX could be used for navigation instead.

Navigation was tested as follows. A taxonomy was learned from a collection of images using the color histogram representation (section 4.1.1). Several (typically, 50 for human subjects) trials of looking up an image in this collection were conducted. In each trial, an image was selected from the collection at random and shown to the human subject. The task was to find this image in the taxonomy. Note that this scenario was only used to ensure that the task has a well-defined answer. In a practical application, the target image would not be known in advance and the user would browse the collection until a 'satisfactory' image is found.

To find the target image, the user started browsing the taxonomy at the root. At each node, all children of that node were displayed (Figure 6), represented by the quilt images. The user could select the most promising (in terms of color similarity) child by clicking on the corresponding quilt image. This moved the user to the corresponding child node. For this child node, all children were displayed again (the possibility to go back up a level was also present), and the process continued until a leaf node was reached. In each leaf, all images in the collection were sorted according to decreasing likelihood of the image in the selected leaf. Note that in each leaf all images in the collection are present, rather than just the images belonging to that leaf. This was done to eliminate backtracking.

Note that sorting images by likelihood involves comparing likelihoods of different images. As mentioned in section 3.1, this comparison is influenced by the number of detections in different images. In our current model, images with fewer detections intrinsically have higher likelihoods (section 3.1). This influence is undesirable. To eliminate it, we ensured that all images had the same number of detections in all our experiments.

The cost of finding the target image is equal to the number of images that had to be examined during search. For an intermediate node, this is the number of quilt images displayed at that node (e. g. 15 in Figure 6), or, equivalently, the number of the node's descendants. For leaf images, the cost is the index of the target image in the sorted collection of images (this corresponds to the time required to locate the targed in this collection). Thus, the more likely the target image is under the chosen path, the less costly the path is.

In an additional set of experiments, a computer algorithm was used to look up an image instead of a human subject. The browsing in these experiments was performed as described above, except that the most promising branch at each step was determined by computing the probability of the query image under each candidate node and selecting the node that gave highest probability. Also, more trials were conducted (typically, one trial for each image in the collection). These experiments were performed mainly to validate and extend the results obtained with human subjects. Note, however, that computer navigation of taxonomies could also have potential applications. For example, it could be used (possibly in combination with hashing) to find duplicate images, or to find images violating copyright (by searching for a given image in the collection of copyright-protected images).

First, we describe computer experiments with the collection of 300 Corel images. Four taxonomies, of depths two, three, four, and five, were built. The average cost of finding each

of the 300 images using each taxonomy was computed. The results are displayed in Figure 7(a) (red bars). These costs were compared to the average cost to locate the target image in an unorganized image collection. If there are N images in the collection, this baseline cost is clearly $N/2$; thus, in our case the baseline cost is 150. Note that any single ordering of these images will incur the same baseline cost for navigation. Therefore, any model (such as LDA) that doesn't form multiple groups (and thus multiple orderings) will perform at baseline level.

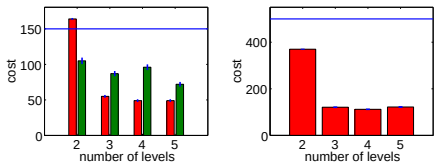
Notice that the two-level taxonomy in this case performs slightly worse than the baseline. The reason is that in this particular case a two-level organization doesn't help navigation; thus, on average the full baseline cost is incurred at the leaves, plus there is the added cost of navigating to a leaf through the taxonomy.

A three-level taxonomy performs much better and in particular significantly outperforms the baseline. This suggests that using a taxonomy is preferable to a flat organization (such as clustering). Using more than three levels in the taxonomy does not improve navigation performance further. This indicates that for the Corel 300 dataset, organization with three levels is sufficient.

For comparison, green bars in Figure 7(a) show the performance of taxonomies randomly sampled from the prior. For each depth value, 20 tree structures were sampled at random from the NCRP prior. The average number of nodes was kept the same as in the learned taxonomy with the same depth. (Taxonomies with number of nodes significantly different from that performed poorer.) The taxonomy structure was fixed, and the ℓ and z variables (i. e., the topics and the categories) were sampled from the posterior for 1000 iterations. The navigation performance of the resulting random taxonomies was averaged. As can be seen, random taxonomies perform significantly worse compared to learned taxonomies. The conclusion is that it is beneficial to tune the tree structure to fit a given collection of images.

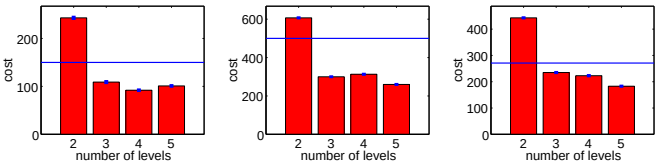
Experiments on an additional image dataset are reported in Figure 7(b). The main conclusion is, again, that relatively deep taxonomies outperform shallower taxonomies. The results are statistically significant ($p < 0.0001$, t-test). Experiments on additional datasets are described in the Supplemental Material.

In addition to the computer experiments above, several experiments with human subjects were performed. Four subjects (three male, one female), three of them not associated with the project, participated in the experiment. The subjects' usual work environment (including monitor, resolution settings, lighting, etc.) was used. The viewing distance and presentation time were not controlled. This is consistent with the proposed application (searching for images in an online collection). The average cost over 50 trials was computed. The results on three image datasets are shown in Figure 8. As can be seen, using relatively deep taxonomies is preferable to flat clustering. The results are statistically significant ($p < 0.0001$, t-test). Overall we found that humans were less efficient than the computer in finding the matching image, indicating that the quilt image representation is perhaps a bit too impoverished for guiding a human effectively.



(a) Corel 300 (b) Corel 1000

Fig. 7. Computer navigation experiments. X axis: number of taxonomy levels used. Y axis: cost (lower is better). Error bars: standard error of the mean. Horizontal blue line: baseline cost (defined as $1/2$ of the number of images in the corresponding dataset). (a): The Corel 300 dataset (section 4.1.1). (b): A subset of 1000 images from the Corel dataset (collections 1000, 103000, 108000, 112000, 118000, 122000, 127000, 132000, 138000, and 143000). In (a), the performance with learned taxonomy structure is shown in red, and the performance with random taxonomy structure is shown in green. In (b), only the performance with learned taxonomy structure is shown.



(a) Corel 300 (b) Corel 1000 (c) Shore

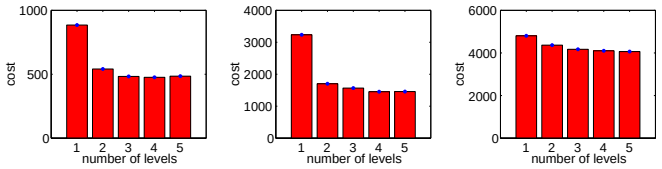
Fig. 8. Human navigation experiments. X axis: number of taxonomy levels used. Y axis: cost (lower is better). Error bars: standard error of the mean. Horizontal line: baseline cost (defined as $1/2$ of the number of images in the corresponding dataset). (a): Corel 300 dataset (section 4.1.1). (b): Corel 1000 dataset (Figure 7). (c): A set of 542 photographs by S. Shore.

4.2.2 Description length of image collections

The defining principle of our taxonomies is that shared information is represented just once at the shared taxonomy nodes. It is therefore natural to measure how well a set of images can be compressed using taxonomies. There are also theoretical arguments for using compression as a measure of learning. Basically, any regularity in the data can be exploited for compression, and it's natural to think of learning as finding regularities in the data. Therefore, in a sense, learning and compression are equivalent [26].

Note that by compression in this section we mean lossless compression of our representation of an image collection. Recall that each image is represented as a bag of words. We compress this representation, and the measure of this compression is usually called 'description length'. Note in particular that we can only reconstruct the bag of words from this compressed representation, not the full original images.

The cost of encoding our image representation is derived in the Supplemental Material. Briefly, the cost consists of two terms. One term measures encoding length of the actual data (detections). A good model would facilitate compression



(a) Corel 300 (b) Corel 1000 (c) 13 scenes

Fig. 9. Description length as a function of the number of levels in the taxonomy. X axis: number of levels. Y axis: description length (lower is better). (a): Corel 300 dataset (section 4.1.1). (b): Corel 1000 dataset (Figure 7). (c): 13 scene categories dataset (section 4.1.2).

by learning topics ϕ_t and categories π_c that allow efficient representation of the detections. For example, encoding a detection from a sparse topic (with high probability on only a few words) is generally less costly compared to a detection from a uniform topic (with roughly equal probability on many words). The reason is that encoding is optimal when the code is designed for the actual distribution of the data (and thus has high probability at the data-cases); therefore, if the distribution of detections is sparse, it is best to encode it using that distribution. In addition, the entropy of a sparse distribution is lower than that of a uniform distribution, implying lower description length. This latter observation also implies that, for example, the number of topics T does not need to be small to achieve efficient encoding. The encoding length will depend on the entropy of distribution over topics and therefore only on the number of topics that are used, rather than on the total number of topics.

The second term controls the model complexity and penalizes overly complex models. For example, a representation in which topics are too sparse (e. g. each topic has non-zero probability on only a single visual word) would require too many topics and thus would be penalized.

Generally, better models achieve lower description length. For intuition, consider a simple example of histogram-based encoding of a set of images. A naive way to perform encoding is to use a single color distribution to represent all pixels in all images. But suppose the images consist of two categories with non-overlapping color distributions (say, ocean images with predominantly blue pixels and forest images with predominantly green pixels). In this case, a more efficient encoding can be achieved by first specifying the category of each image, and then using a category-specific code to represent all pixels in that image. Category-specific codes are more efficient in this case since each code needs to represent only a fraction of all available colors. Specifying the category label does introduce some overhead, but it is generally negligible since only one category label is sufficient for all pixels in a given image. More elaborate models (such as taxonomies) will achieve even better description length if they fit the data well.

We have performed compression experiments on several datasets. Figure 9 describes the datasets and shows the results. As can be seen, the results are generally consistent with the previous navigation experiments, in that a relatively deep taxonomy outperforms shallower taxonomies. Description lengths achieved by a depth-1 taxonomy are also included for comparison. Experiments on additional datasets are described in the Supplemental Material.

Note that the relative improvements in going to deeper taxonomies are smaller than in the case of navigation. A possible explanation is that although images share a relatively small amount of information (this is indicated by the small decrease in description length), even this amount is sufficient to organize images hierarchically and achieve a significant improvement in navigation. As an analogy, recall that in natural images, gradients are very sparse, but provide the majority of information about the image. We stress again that relatively deep taxonomies outperform flat organization and shallow taxonomies in all of our experiments.

4.2.3 Comparison to hierarchical clustering

Both the TAX model itself (section 3.1) and the inference procedure in TAX (section 3.2) are quite complex. It is interesting to compare TAX to a simpler model. We implemented two versions of hierarchical clustering for comparison. These are described next.

In the first method, the entire image collection was first split into two groups using multinomial k-means clustering. The cluster 'centroids' became the first two topics. Each of these groups was further split in two by repeated application of multinomial k-means. This process continued until a binary tree of prescribed depth ($L = 4$ in our experiments) was obtained. The resulting tree was used for navigation as follows. Each target image started at the root of the tree. Its likelihood was evaluated under the topics of both of the root's descendants, and the most probable descendant was selected. The navigation continued in this manner until a leaf was reached. As in TAX, each leaf contained all images in the collection, sorted by their likelihood in that leaf. The navigation performance of this model on the Corel 300 dataset was 58 ± 3 (average $\pm$ standard error of the mean), compared with 49 ± 2 achieved by TAX (lower values are better). On the Corel 500 dataset, hierarchical clustering achieved the cost of 89 ± 3, compared with 70 ± 3 achieved by TAX. Both results are statistically significant (t test, $p < 0.01$).

In a second hierarchical clustering model, a set of topics is first learned using LDA. This gives the topic indices $z_{i,d}$ for all detections. Subsequently, a taxonomy is learned using these pre-defined topics. This is performed by running a modified TAX inference procedure in which the $z_{i,d}$ variables are not resampled, but rather are kept fixed to their previously learned values. All other variables (including the paths and the $l_{i,d}$'s) are resampled as normal. We evaluated the hold-out set perplexity (e. g. [4]) the resulting model achieves, as well as its navigation performance. On the Corel 300 dataset, the resulting model achieves perplexity of 3.51, compared to 3.16 achieved by regular TAX (lower perplexity values indicate better performance). The navigation performance of hierarchical clustering was 91, compared to 49 achieved by regular TAX.

The conclusion is that the added compexity of TAX is justified by the significantly increased performance it offers.

4.3 Beyond taxonomies

While a taxonomy is a natural and efficient way to organize images, it is clear that real-world images need not have strict taxonomic structure and may be organized in a more complex manner. In this section, we explore some evidence for such more complex organization.

Recall that in TAX, the organization is defined by how information is shared among multiple images (section 3). Although alternative organizing principles may exist, considering those is beyond the scope of this paper. Here, we limit ourselves to exploring structures arising from information sharing.

Below, we present two experiments which suggest that extending our model beyond taxonomies may be necessary. Note that these experiments are performed on the Corel 300 dataset (section 4.1.1). We have already shown above that the taxonomy learned from this dataset facilitates representation and improves performance of two visual tasks. This indicates that even though a taxonomy may not be the optimal representation for a given image collection, it is nevertheless a useful representation that has some benefits compared to a flat model.

4.3.1 Multiple variants of taxonomies

Using Gibbs sampling for inference allows us to recover several modes of the underlying distribution. We re-ran the sampler 15 times with different random initializations on the Corel 300 dataset. Although there were small differences between all 15 resulting taxonomies, only four of them were substantially different. (This was judged informally, by the authors visually examining all 15 taxonomies. It is not critical for our purposes to perform a more objective evaluation, because even the four remaining taxonomies are quite similar.) These four taxonomy variants are shown in Figure 10.

Some of the nodes in Figure 10 are labeled as follows. The leading number (e. g., '1' in label '1A') is the variant number (there are four variants total). The letter in the label denotes a specific node in a variant. The letter is uppercase for second-level nodes and lowercase for third-level nodes. Other than that, the letters are in no particular order.

As can be seen, many nodes are similar between the four taxonomy variants, even though the overall structure is somewhat different. For example, note the similarity between the following groups of nodes: [1A, 3A]; [1C, 2C, 3C]; [1D, 2D, 3D, 4D]; [1c, 2c, 3c, 4c]; [1a, 2a, 3a, 4a]; [1b, 3b]; [1d, 4d]; [1e, 2e, 3e]; [1h, 3h]; [2k, 4k]; [2j, 3j]. Similarities between both high-level nodes (e. g., [1A, 3A]) and lower-level nodes (e. g., [1c, 2c, 3c, 4c]) are present. Similarity between two high-level nodes can be complete (when all their children also match, as between 1C and 2C) or partial (when only some of the children match, as between 1A and 3A). In some cases, the structure of a low-level node is preserved between two variants, but the node is grouped with different siblings. This can be thought of as the node choosing a different parent. For an example, consider nodes [1h, 3h].

Two conclusions can be made from these results. First, the taxonomies are quite stable, and the sampler is able to reliably recover the main components of the taxonomy. Second, no single taxonomy structure is best. There are multiple ways to organize the individual nodes into the taxonomy, and multiple ways to share the common information. It is desirable to exploit all of these simultaneously, and a single static taxonomy is insufficient for this purpose. Drawing multiple samples from the model (as done here) is one way to approach this issue. Alternatively, a model more complicated than a taxonomy may be used to represent all of the possible modes simultaneously.

4.3.2 Secondary structure

As mentioned above, our model organizes images according to how information is shared among them. This shared information is represented by topics common to multiple images. In a strict taxonomy, topics would be shared only via the taxonomic structure: the common topics would be represented at common top-level nodes, and images using these topics would share these common nodes. A particularly interesting feature of TAX in this context is that TAX can represent both a strict taxonomy in the sense described above, and a more general structure (a cyclic graph). The reason is that all topics are available at all nodes; as a result, any two nodes, even those not in a hierarchical relationship, can potentially share topics. Such sharing increases the statistical power of the model. In addition, it helps discover the natural non-hierarchical relationships between image categories.

Figure 1 shows the secondary sharing structure in the Corel 300 taxonomy. The black edges represent the original taxonomy structure. The red edges connect nodes that are not in a hierarchical relationship with each other, but nevertheless share a topic. Two nodes are said to share a topic when both assign a substantial probability (above 0.1 in our experiments) to that topic. The bottom part of Figure 1 shows the shared topics. These topics are distributions over visual words which represent color. For each topic, five visual words most probable in that topic are shown in the order of decreasing probability. (Some topics, e. g. topic 39, assign non-zero probability to less than five words; fewer than five words are shown in such cases.) The colored squares at the bottom represent the visual words. The height of the vertical bar above each word is proportional to the word's frequency in the topic. As can be seen, the primary structure of this dataset is well approximated by a taxonomy, but there is also substantial secondary structure which turns the taxonomy into a cyclic graph.

Note that it would be difficult to encode this graph by a strict taxonomy. One possibility would be to push each shared topic up to the lowest common ancestor. This, however, would decrease the quality of the taxonomy. For example, nodes 1 and 4 share topic 39. If that topic were pushed into the common node (the root node in this case), then the representation of all images in the taxonomy would include that topic. In practice, many images (nodes 2, 3) don't use that topic, so including it in their representation would be wasteful.

The general conclusion is that a strict taxonomy cannot fully represent the sharing structure in an image collection. Note again that our model could in principle learn a strict taxonomy if a strict taxonomy were sufficient. The fact that a cyclic graph is obtained instead strongly suggests that a more elaborate representation for image collections may be necessary.

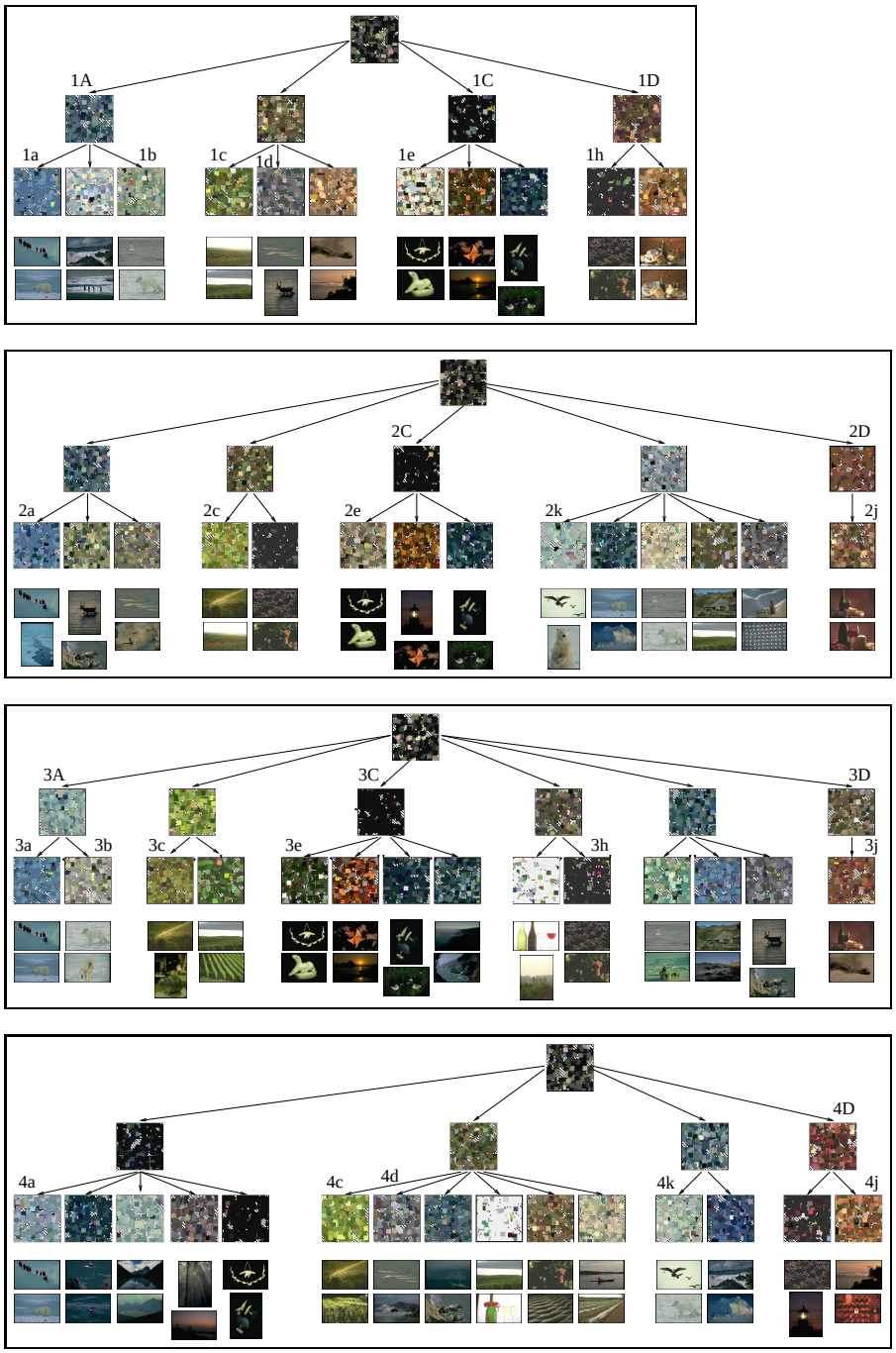


Fig. 10. The four stable variants of taxonomies we found by running TAX 15 times on the Corel 300 dataset. The display is similar to Figure 1, except that only two (rather than four) most probable images are shown below each leaf. As can be seen, many nodes are similar between these variants, even though the overall structure is somewhat different. See section 4.3.1 for more details.

5 Discussion

We experimented with organizing image collections into taxonomies in an unsupervised manner. The general conclusions are that taxonomies may be learned in an unsupervised manner and that such organization significantly improves representation and boosts performance when searching an image collection. These results hold even in cases where a representation more sophisticated than a strict taxonomy may be involved; therefore, learning taxonomies may be useful even if the correct structure for a given collection is a priori unknown.

A non-parametric Bayesian model called TAX was used for learning the taxonomies. It is based on the Nested Chinese Restaurant Process (NCRP) [17]. TAX can learn taxonomies for a wide variety of datasets. In addition, TAX can learn and implicitly represent the structure of general cyclic graphs.

One of the main limitations of TAX is the speed of training. For example, with 10,000 training images, training took about 7 days. Significant progress in computational efficiency is clearly needed. Preliminary results indicate that methods based on grouping similar variables [27] can significantly improve the speed and memory requirements of inference.

Speeding up learning is also crucial for applications such as internet search. Consider a scenario where the user types a query word (such as 'cow'), and a set of matching images is retrieved. Organizing these images hierarchically would be very useful, but the process needs to be very efficient in order not to slow down the search. In the future, we plan to explore using methods analogous to cross-generalization [3], [2] to efficiently modify an existing taxonomy to match a new query.

Several results in section 4.3 indicate that a structure more elaborate than a taxonomy may be needed for some image collections. Learning and exploiting such structures is the main goal for future research.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. 0447903, No. 0535278 and IIS-0535292, and by ONR MURI grant 00014-06-1-0734. An early version of this paper appeared in [18]. The authors would like to thank Marco Andreetto for useful suggestions.

References

[1] I. Biederman, "Visual object recognition," in *An Invitation to Cognitive Science*, 2nd ed., S. F. Kosslyn and D. N. Osherson, Eds. MIT Press, 1995, vol. 2, pp. 121–165.
[2] L. Fei-Fei, R. Fergus, and P. Perona, "A Bayesian approach to unsupervised one-shot learning of object categories," in *ICCV*, 2003.
[3] E. Bart and S. Ullman, "Cross-generalization: learning novel classes from a single example by feature replacement," in *CVPR*, 2005.
[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J Amer Stat Assoc*, 2006.
[5] Y. Amit, D. Geman, and X. Fan, "A coarse-to-fine strategy for multiclass shape detection," *PAMI*, vol. 28, pp. 1606–1621, 2004.
[6] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *CVPR*, 2008.
[7] A. Zweig and D. Weinshall, "Exploiting object hierarchy: Combining models from different category levels," in *ICCV*, 2007.
[8] X. He and R. Zemel, "Latent topic random fields: Learning using a taxonomy of labels," in *CVPR*, 2008.
[9] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* John Wiley & Sons, 2001.
[10] M. Andreetto, L. Zelnik-Manor, and P. Perona, "Non-parametric probabilistic image segmentation," in *ICCV*, 2007.
[11] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing visual scenes using transformed Dirichlet processes," in *NIPS*, 2005.
[12] E. Miller, N. Matsakis, and P. Viola, "Learning from one example through shared densities on transforms," in *CVPR*, 2000, pp. 464–471.
[13] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *CVPR*, 2004.
[14] M. Fink, S. Shalev-Shwartz, Y. Singer, and S. Ullman, "Online multiclass learning by interclass hypothesis sharing," in *ICML*, 2006.
[15] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan, "Matching words and pictures," *JMLR*, vol. 3, pp. 1107–1135, 2003.
[16] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006.
[17] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," in *NIPS*, 2004.
[18] E. Bart, I. Porteous, P. Perona, and M. Welling, "Unsupervised learning of visual taxonomies," in *CVPR*, 2008.
[19] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised discovery of visual object class hierarchies," in *CVPR*, 2008.
[20] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005.
[21] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.
[22] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their localization in images," in *ICCV*, 2005, pp. 370–377.
[23] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *J Amer Stat Assoc*, vol. 90, no. 430, pp. 577–588, 1995.
[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
[25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.
[26] P. Grunwald, "A tutorial introduction to the minimum description length principle," arXiv:math.ST/0406077.
[27] E. Bart, "Speeding up gibbs sampling by variable grouping," in *Proceedings of NIPS Workshop on Applications for Topic Models: Text and Beyond*, 2009.

Evgeniy Bart received PhD degree in computer science from the Weizmann Institute in 2004. He participated in the IMA annual imaging program in 2004–2005, and currently has a post-doctoral position at Caltech. His research interests include computer vision and machine learning.

Max Welling has received his PhD in theoretical physics. After this he has switched fields to computer vision and machine learning. He is currently appointed in two departments, computer science and statistics, at UC Irvine.

Pietro Perona is the Allen E. Puckett Professor of Electrical Engineering at Caltech. He has contributed to the theory of partial differential equations for image processing and boundary formation, and to modeling the early visual system's function. He is currently interested in visual categories and visual recognition.