Probabilistic Visual Concept Trees

Lexing Xie[†], Rong Yan[‡], Jelena Tešić[*], Apostol Natsev[†], John R. Smith [†]
[†]IBM T. J. Watson Research Center, [‡]Facebook Inc., [*]Mayachitra Inc.

ABSTRACT

This paper presents probabilistic visual concept trees, a model for large visual semantic taxonomy structures and its use in visual concept detection. Organizing visual semantic knowledge systematically is one of the key challenges towards large-scale concept detection, and one that is complementary to optimizing visual classification for individual concepts. Semantic concepts have traditionally been treated as isolated nodes, a densely-connected web, or a tree. Our analysis shows that none of these models are sufficient in modeling the typical relationships on a real-world visual taxonomy, and these relationships belong to three broad categories – semantic, appearance and statistics. We propose probabilistic visual concept trees for modeling a taxonomy forest with observation uncertainty. As a Bayesian network with parameter constraints, this model is flexible enough to account for the key assumptions in all three types of taxonomy relations, yet it is robust enough to accommodate expansion or deletion in a taxonomy. Our evaluation results on a large web image dataset show that the classification accuracy has considerably improved upon baselines without, or with only a subset of concept relationships.

Categories and Subject Descriptors

H.3.3 [**Information Systems Applications**]: Information Search and Retrieval

General Terms

Algorithms, Design, Experimentation

1. INTRODUCTION

Visual concept detection is an important problem that has received significant recent attention. The research community has accumulated significant collective knowledge on this problem, along with the increasing performance in large public benchmarks [7, 8]. There are two main challenges for concept detection to real-world scale. The first is scaling up with the amount of data, this include devising learning algorithms to learn from, and performs robustly to large amounts of diverse media content. The second one is scaling up with semantics, that is, designing learning mechanisms that scales gracefully to a large number of visual semantics, given that most current research in visual recognition tagets a few dozen concepts [7, 8] among tens of thousands in the real-world. This paper is concerned with the second challenge.

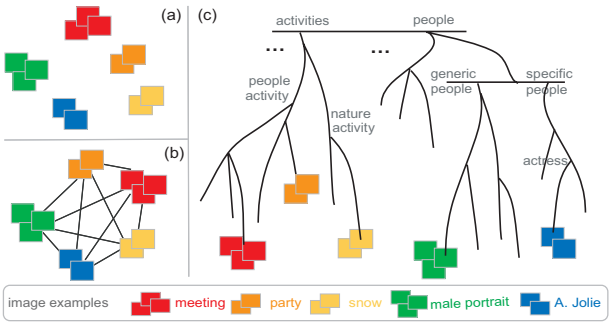

Figure 1: Three views of an example visual concept collection: (a) isolated dots; (b) concept web; (c) concept forest.

Systematic organization is key for acquiring large amounts of knowledge, and this applies to the learning process of human and machines alike. Taxonomy, as the practice and science of classification [9], is a natural tool for organizing facts and entities. Taxonomies are widely used in areas ranging from biology, medicine, military operations to web design. Visual semantic taxonomy is calling for attention once visual recognition proceeds beyond a few specific categories such as faces, human and cars. Many existing work optimizes the detection of each specific visual concepts by making independent binary decisions, this is equivalent to treating a collection of concepts as isolated dots. A number of prior investigations have modeled pair-wise inter-concept relationships [2, 10], equivalent to densely connecting the concepts into a web, or a tree-structured hierarchy with mutually exclusive relationship [1]. While tree structures are natural for concept organization, these trees are typically small, and it is too rigid to take into account that an image has multiple labels based on different aspects of semantics, e.g., a portrait of a famous actress can be described as *female face*, *actress*, *one person*, each based on the gender, role and number of person, respectively.

We propose a taxonomy representation and inference structure that can take into account such relationships. We start by presenting three salient relationships about real-world visual taxonomies in large scale, namely: concept semantics, image appearance and data

MM'10, October 25–29, 2010, Firenze, Italy.

statistics. We present a multi-faceted concept forest structure that conceptualize these relationships, including the parent-children relationship, mutual exclusion relationship, as well as the multiple aspects labeling such as in the female actress portrait. Fig. 1 shows a comparison between this structure, and the two prior alternatives: concept dot and concept web. We also propose *Visual Concept Trees* that computationally encodes a multi-faceted concept forest under observation uncertainty. This model evolves from tree-structured Bayes Net, and is designed to extend a few predecessors such as the Bayes Net and tree-structured taxonomy. It is learned from observations and performs inference with the junction tree algorithm. We evaluate this model on a large web image collection consisting of thousands of images and hundreds of visual concepts. We use the concept tree structure to post-filter binary classifier outputs, and observed up to 0.4 improvement in classification accuracy, and behaves robustly with inaccurate concept priors.

2. TAXONOMIES FOR VISUAL SEMANTICS

In order to design taxonomy-aware concept models, we start by examining the different types of concept relationships in images and videos. We present a list of relationships found useful based on both generic semantic knowledge in WordNet [5], and the practices in building large taxonomies. In this list, there are two semantic-driven relationships that are broadly applicable and robust to taxonomy variations, two appearance-driven relationships frequently seen in image and video collections, and two statistics-driven aspects accounting for uncertainty in data collections.

2.1 Semantic-driven relations

Being "visually detectable" is one of the primary criteria for semantic modeling in images and videos. These include concrete nouns [1], and a subset of verbs that can be captured in a visual scene or translated to the corresponding noun, e.g., *protest*, *concert*, *walking*. A generic semantic lexicon such as the WordNet [5] has more than a dozen relations among nouns, verbs, adjectives and adverbs. We specifically choose two types of relations for visual taxonomy: (1) Parent-children relationships. This maps to *hypernyms* and *hyponyms* in WordNet terms, i.e. every instance of concept A is a (kind of) concept B. An *apple* is a *fruit*, and *walking* is a kind of *movement* for instance. (2) Mutual exclusion. This maps to *coordinated terms* in WordNet which share a common *hypernym*. *Apple*, *organge* and *watermelon*, *walking* and *jogging* are examples of mutually exclusive concept sets.

We choose these two relations as they are applicable to concrete nouns and verbs, and they are robust to typical visual appearance variations. For example, the "part-of" relationship (*holonym* and *meronym* in WordNet) is often violated in visual appearances, as we often see photographs of a *window* without the *building* it is attached to, or closeup shots of a *tree* without its *trunk* visible. Finally, these two relations are can be identified in limited context when working on a single piece of image or video segment. The "entailment" relationship (A is a result of B), on the other hand, requires more than one image or video to be analyzed in the order of causality and temporal precedence, i.e. we do not have knowledge of a soccer match just by seeing the award ceremony that followed.

[1]A concrete noun refers to objects and substances, including people and animals, that exist physically, e.g., *chair, apple, clock*. An abstract noun refers to states, events, concepts, feelings, qualities, etc., that have no physical existence. e.g., *freedom, happiness, music*.

2.2 Appearance-driven relations

Traditional wisdom has it that "a picture is worth a thousand words". On image and video datasets for recognition, this means that an image is often associated with multiple labels, such as *park, party, crowd, trees*. In addition, there is often more than one way that we can use to further classify a concept. For example pictures containing *people* can be further classified according to the number of people, their age, their poses and actions, or their occupations.

2.3 Statistics-driven relations

The goal of automatic recognition is to tag visual concepts from noisy observations, including low-level features computed directly from images, or predictions from mid-level semantic classifiers. There are two main types of uncertainties in the observations: (1) Relationships between concepts and observations, such as combining two classifier with 65% and 60% accuracies would help infer the true labels more accurately than either of the two. (2) Statistical relationships among observations, possibly upon different concepts that do not have a clearly prescribed relationship in the taxonomy. Such as seeing *beach* and *palm trees* in a picture enhances the likelihood of also seeing *sky*. These relationships have been shown to be useful to help classification [2, 10].

2.4 Comments on large lexicons

Another reason for choosing the above relations is to account for the flexibility and fluidity of a large-scale visual taxonomy. Unlike classifying living species, there is no Linnaean taxonomy of visual semantics – people's view of what worth classifying and how to classify them tend to change with respect to application domains, data collections, an the evolving knowledge about the semantics. Moreover, taxonomies grow over time as new concepts and new categories evolve, such as *Wii* as a new video game system as a sibling to *XBox*. Note that semantic relations are natural constraints preserved through the changes in taxonomy. Being parent-children (including grandparent) or mutual exclusive still holds true even after new nodes or branches are added. The appearance and statistical relations are also invariant to concept insertions or revisions since the are essentially grounded in the underlying data domain.

3. PROBABILISTIC CONCEPT TREES

Using the semantic and data-driven relationship as guidelines, we introduce Probabilistic Concept Tree to encode them via series of models.

3.1 Earlier models

The naive Bayes model is a simple model for the two statistical-driven relationships (Sec. 2.3). This is done by factoring the joint class-probabilities into the product of multiple independent conditional probabilities given a concept class label, as shown in Eq. 1. It estimates the class-conditional probabilities from observations x, and finds the most likely class based on the Bayesian rule. Fig. 2(a) show the form of model that has been effectively used in prior work [10] in which the concept labels y are considered as binary, e.g. *apple*, *not-apple*. A simple extension to the binary naive Bayes model is to consider multi-valued labels that are mutually exclusive (e.g., $y \in \{$ *apple*, *orange*, *peach*, $\dots\}$), thus also capturing semantic mutual exclusion. This model has the same graphical form (Fig. 2(b)) as its binary variant, except that the posterior probabilities among the sibling concepts become related.

$$P(y|x_{1:M}) \quad \propto \quad P(y) \prod_{i=1:M} P(x_i|y_c) \qquad (1)$$

Note that the hierarchical parent-children relationship in concept semantics is notably missing from the naive Bayes models. Un-

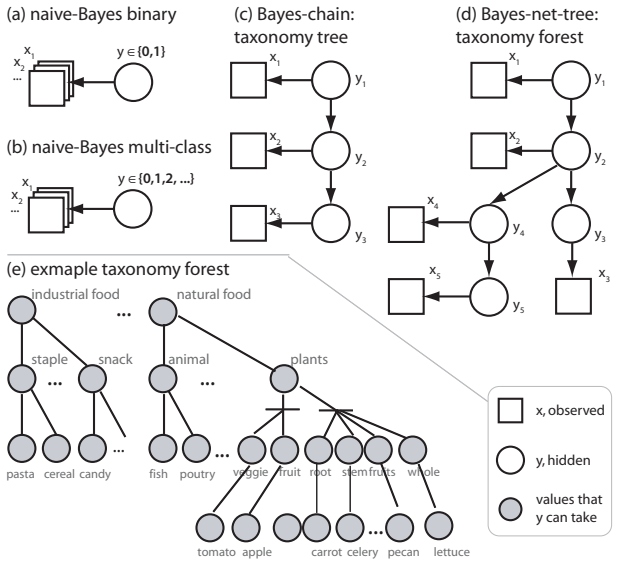

Figure 2: Overview of various taxonomy models, see Sec. 3 for descriptions.

der this relationship, we can organized semantic concepts can be organized into a tree, with parent nodes pointing to the children nodes, and the conditional probabilities control their membership probabilities. As shown in Fig. 2(c), the chain of hidden variables y can represent a tree structure in its state-space, with each y node taking multiple possible values (e.g. *apple*,*orange*,...), and constraints in the conditional probabilities that set the values between non-parent-children node pairs to zero (e.g., $P(fish|plant) = 0$).

3.2 Construction of Probabilistic Concept Tree

We notice that multiple classifications can be represented by multiple decision variables y simultaneously taking values on different state spaces. In (Fig. 2(e)) for example *food from a plant* can be a *fruit* or a *vegetable*, and at the same time it can also come from the *root*, *leaves* or *stem* of the plant. We also notice that a tree-structured taxonomy is naturally recursive, i.e. concepts that belong to *fruit* or those belong to *vegetable* can be organized into a tree or forest themselves. In concept state-space these two designs translate to a taxonomy "forest", and in graphical model this can be manifested with parallel branches in the hidden states. This leads to the tree-structured Bayesian network, dubbed *Probabilistic Concept Tree*, shown in Fig. 2(d). The previous models account for semantic-driven and statistics-driven relations, and this adds the appearance-driven relations, as it allows multiple labels in the same image, and models the loose correlation among them.

We can use a recursive process to construct such a tree from an existing taxonomy forest by walking the forest in the following few steps: (1) Add a node for its root. e.g., $y_1 \in \{natural\ food, industrial\text{-}food, \dots\}$. (2) Add a node for each set of children of the same generation e.g., $y_2 \in \{staple, snack, plants, \dots\}$. (3) Add a branch for each parallel subtree. i.e., $y_3 \in \{veggie, fruit, \dots\}$ and $y_4 \in \{root, stem, \dots\}$ (4) Repeat steps (2) and (3) until all states are added to the tree. This particular instance corresponds to the example taxonomy forest in Fig. 2(e).

The parameters of a probabilistic concept trees include three parts: the "emission" probabilities $p(x|y)$ of seeing the observations x given state variable y, the hierarchical conditional probabilities $P(y_i|y_{Pa_i})$ between a state variable y_i and its parent variable y_{Pa_i}, as well as the prior $P(y_{root})$ on values that the root node can take. Multi-variate Gaussian conditional probabilities are used for the real-valued observations x_i and the corresponding state y_i; tabular conditionals are used between pairs of states $P(y_i|y_{Pa_i})$. Note that the concept tree construction requires block-wise assignment of conditional probabilities based on parent-children relationships in the taxonomy. For instance, in the second level of Fig. 2(d) and (e) the conditionals need to be set such that

$$\sum\nolimits_{y_2 \in \{animal, plants, \dots\}} P(y_2 \mid y_1 = natural\ food) = 1\ ;$$
$$\forall y_2 \in \{staple, snack, \dots\},\ P(y_2 \mid y_1 = natural\ food) = 0\ .$$

According to these network and its parameter constraints, we can write out the joint probability of all observations and hidden states in a Probabilistic Concept Tree in Bayes network notation as in Eq. 2.

$$P(x_{1:M}, y_{1:M}) = P(y_{root}) \prod_{i=2}^{M} P(y_i|y_{Pa_i}) \prod_{i=1}^{M} P(x_i|y_i) \qquad (2)$$

We use expectation-maximization (EM) to estimate model parameters from training images, and use the junction tree algorithm [3] to estimate posterior probabilities of $P(y_i|x)$. The model inference is carried out with the block-wise constraints in the conditional probability tables. The inference on a probabilistic concept tree is efficient: linear in the number of nodes and quadratic in the size of the state-space. This can be implemented using Bayes Network tools such as the BNT [6]. Due to space constraints we omit further details of the model and its inference steps.

4. EXPERIMENTS

We evaluate Probabilistic Concept Trees on a web image collection containing 60,200 images collected from different photo sharing sites and internet search engines. The images are of diverse semantics and are manually filed into a taxonomy of 222 concepts in total. The taxonomy is manually designed, organized into six top-level categories, in a hierarchical forest similar to the illustration in Fig.1(c). The six top-level facets are *activities, domain, objects, people, setting* and *image type*, each contain $10 \sim 50$ concepts, with a depth of $3 \sim 7$. Each category is modeled by a Bayes concept tree of $4 \sim 16$ nodes. Each image in our dataset are filed into one or more leaf nodes among the six concept trees. We split this dataset and use $2/3$ for training, $1/3$ for testing.

For each of the 222 categories we first train an ensemble of Support Vector Machines (SVM) on positive and negative examples of each concept. The training images are taken from a separate collection of about 240K web images, and negative examples for each category are taken from the rest of the taxonomy forest using the semantic relationships to infer mutual exclusion. Details about training and testing our semantic concepts can be found in [4]. These serves as the classification baselines and input to the concept tree models. The learning and inference of Probabilistic Concept Trees are efficient. The training and testing of all models would finish within a few hours on a single CPU with a matlab implementaion.

We compare concept label prediction among the following five methods: (1) Original classifier score (abbrv. *orig* or *o*). (2) Binary Naive-Bayes binary (*nb-bin* or *b*), as shown in Fig. 2(a). Its input are SVM scores, the output are binary confidence scores for each concept. (3) Multi-class naive-Bayes model (*nb-multi* or *m*), as shown in Fig. 2(b), the input are SVM scores, the output are posterior probabilities of sibling concepts and maximum a posteriori class labels. (4) Bayes net concept tree (*bnet*), shown in Fig. 2(d). It explicitly models parent-children, mutual exclusion and multi-faceted concept taxonomy. This became equivalent to the Bayes

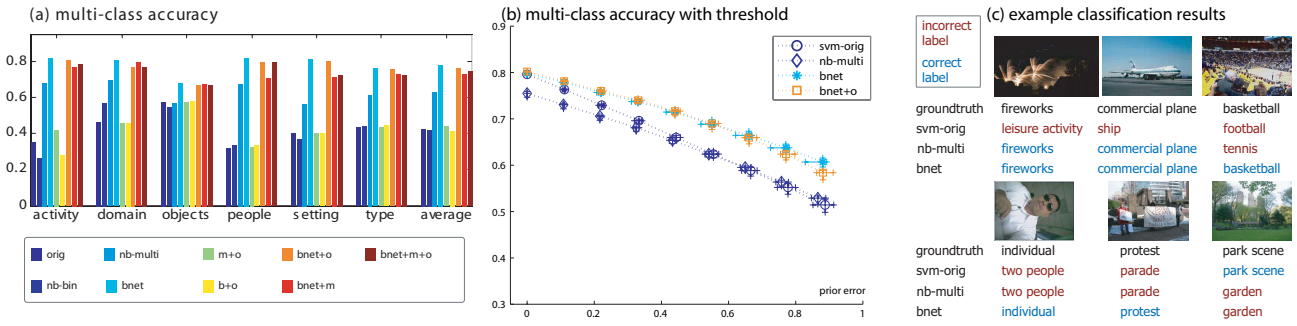

Figure 3: Result summary of different detection methods.

chain structure Fig. 2(c) when the underlying taxonomy state space is one single tree. (5) Various fusion models (*b+o, m+o, bnet+o, bnet+m, bnet+m+o*). We linearly combine the posterior probabilities with equal weights for each target concept. The SVM scores in o are converted to posterior probabilities with the logistic function before fusion.

We measure multi-class classification accuracy among the *sibling* concepts at the same depth of a concept tree. For instance, at depth 2 in Figure 2(e) we will have concepts *staple, snack, animal, plants, none-of-the-above*, and classification accuracies are measured as the fraction of correctly labeled images over all images with known labels with respect to the current tree.

Fig 3(a) compares the classification accuracy of the different models, shown there is the average accuracy over each of the six concept trees as well as the mean over all six. We can see that *nb-multi* and *bnet* are clearly better suited for multi-class classification. This performance is preserved if we combine posteriors from Bayes nets with binary classification scores (*bnet+**), and apparently not so if only *nb-multi* was combined (*m+o*). We also noticed that the performance gain is notably larger with Bayes net when the underlying taxonomy forests has more branching in the network structure (i.e. more legitimate labels per image), such as *people* and *activities*, than those very close to a tree structure, such as *objects*.

One of the reasons why inherently binary classifiers did not perform well is the lack of knowledge on the concept priors $P(y)$. In Fig 3(b) we experiment with the sensitivity of the classification performance to priors with noise. We rank the model posteriors for each dimension (sibling concept) and threshold at a $\tilde{p} = (1-\alpha)p_0 + \alpha u$, where p_0 is prior probabilities estimated from training data, u is prior noise sampled from a uniform Dirichlet distribution, and α is a weight factor ranging from 0.1 to 1.0. We plot the average classification accuracy over all concepts versus the amount of noise in the prior, and the results show that although *bnet* and the *original svm* models are almost on par when the knowledge for priors are correct, but SVM is much more sensitive to deviations from the correct prior.

Fig. 3(c) shows example classification results of the different models. For the *fireworks* and *commercial plane* images on the top, labels are being corrected taking into account the scores of their sibling concepts that are mutually exclusive. For the *basketball*, *individual*, and *protest* images, the labels are corrected only after propagating and re-weighting the observations with information from parent nodes. For the last image of *park scene*, the automatic labels *garden* is visually quite sensible and may even suggest that the related taxonomy may be expanded.

These results demonstrate that Probabilistic Visual Concept Tress represent a class of models effective for encoding hierarchical, mutually exclusive, and multi-faceted concept relationships under uncertainty. Multi-class classification performance is significantly improved. Moreover, the resulting concept scores are more robust to imperfect parameters such as deviations from the prior, as the concept parents and siblings does help produce correct labels in a few notable examples.

5. CONCLUSION

We presented probabilistic concept trees, a novel representation and inference model for large semantic visual taxonomy. This model is distinct in that it accounts for two robust semantic relationships (parent-children and mutual exclusion) as well as the appearance-driven and statistics-driven relations in a visual taxonomy. We derived the parametrization and inference of the model as a special case of Bayesian network. We have observed significant improvement in classification accuracy on a large collection of web images. Future work can include automatic learning of the taxonomy forest structure from data, extensions to discriminative relationships, adding spatial-temporal compositions about event concepts. This model can also potentially be used for concept suggestion in taxonomy design and data annotation.

6. REFERENCES

[1] X. He and R. Zemel. Latent topic random fields: Learning using a taxonomy of labels. In *CVPR*, 2008.

[2] W. Jiang, S.-F. Chang, and A. C. Loui. Kernel sharing with joint boosting for multi-class concept detection. In *IEEE CVPR Workshop*, Minneapolis, Minnesota, 2007.

[3] M. I. Jordan. *Learning in graphical models*. Kluwer, 1998.

[4] M. Campbell et. al. IBM research TRECVID-2007 video retrieval system. In *NIST TRECVID Workshop*, 2007.

[5] G. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[6] K. Murphy. The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, 33(2):1024–1034, 2001.

[7] PASCAL. The VOC challenge, 2005–2008. http://pascallin.ecs.soton.ac.uk/challenges/VOC/.

[8] The National Institute of Standards and Technology (NIST). TREC video retrieval evaluation, 2001–2008. http://www-nlpir.nist.gov/projects/trecvid/.

[9] Wikipedia. Taxonomy. http://en.wikipedia.org/wiki/Taxonomy.

[10] L. Xie, R. Yan, and J. Yang. Multi-concept learning with large-scale multimedia lexicons. In *IEEE ICIP*, 2008.