Layered Object Categorization

Lei Yang[1,2] Jie Yang[2] Nanning Zheng[1] Hong Cheng[1,2]
[1]*Institute of Artificial Intelligence and Robotics, Xian Jiaotong University, P.R.China*
[2]*Human Computer Interaction Institute, Carnegie Mellon University, USA*
{*ylangei, jie.yang, hongc*}*@cs.cmu.edu, nnzheng@mail.xjtu.edu.cn*

Abstract

In this paper, we propose a novel framework of object categorization, namely layered object categorization, which takes advantage of hierarchical category information and performs object categorization at different levels. The proposed hierarchical structure of object categories is built bottom-up and top-down simultaneously accordingly to cognitive rules. First, part-based models are learnt to evaluate structure similarities at the basic level and objects are divided into basic categories. Then the decision cues for object categorization at different layers are optimally selected. Prior knowledge about inter-category relationships is utilized to infer objects' higher inclusive concept labels, while the most discriminative visual details of each category at the lower specific levels are selected automatically. We evaluate the proposed method with a hierarchical database and show promising results. The layered object categorization provides an efficient way for dynamically adapting the object categorization results to different applications.

1. Introduction

Object categorization is a process of assigning a specific object to a certain category, which is also referred as 'generic object recognition' in contrast to 'specific object recognition'. However, towards specific object instance, with various generic degrees, categories are organized in hierarchies, e.g., food-sandwich-burger-beef burger. Human cognition relies on such a hierarchical representation of objects in the process of perceiving and understanding, and can perform object categorization at different levels of the hierarchy. Thus, in terms of performance, humans do much better in categorization than machines.

Until recently, some researchers started to study the multilevel property of categories for automatic object categorization. Paper [8] investigated whether a carefully designed visual vocabulary could support more specific object categorization. In paper [9], authors combined two or more classifiers built at different levels in the object hierarchical tree to do classifications at the most specific level. In paper [1], a part model of the 'basic level' category was used to form a vector representation for the subordinate class recognition. However, these researchers aimed to develop efficient classifiers for object categorization at a specific subcategory level without extracting multi-layered category information. On the other hand, a few researchers attempted to do multilevel categorization ([3][6][7]). In paper [3], the authors proposed a system to address generic class recognition and specific object recognition in the same framework. Paper [7] used lexical semantic networks to build up the semantic category hierarchies. In both papers, the hierarchical classifiers are learnt in a totally top-down manner assuming that objects at all levels could be divided into categories by the same appearance cues. In paper [6], the hierarchical category structure was not built in a semantic way and the optimal multi-cues for the decision tree were manually set.

In this paper, we propose a framework for object categorization at multiple different levels. Based on the previous cognition researches [1], there is an intermediate basic level in the object hierarchy, which should be learnt first. This basic level seems related to the structure of the objects. Higher inclusive levels are more often described by their functions (e.g., food), while lower specific levels are often described by the detailed attributes (e.g., beef burger and chicken burger, etc.). Thus the proposed hierarchical structure of object categories is built bottom-up and top-down simultaneously accordingly to these cognitive rules. First, part-based models are learnt to evaluate structure similarities at the basic level and objects are divided into basic categories. Then the decision cues for object categorization at different layers are optimally selected. Prior knowledge about inter-category relationships is utilized to in-

fer objects' higher inclusive concept labels, while the most discriminative visual details of each category at the lower specific levels are selected automatically. An attribute model combining multiple visual cues is utilized to determine the most discriminative semantic details of categories at each lower level automatically.

2. Part-based models

To distinguish the structure differences of categories at the basic level, we use the relational object class model from [2] due to its computational efficiency and competitive recognition results. The object model is with P parts, where for each part, its appearance, location, and scale are modeled.

Given a learnt model for a certain category and a new image, the likelihood ratio test function is approximated by

$$f(I) = \max_{c} \sum_{k=1}^{p} \max_{x \in F(I)} \log p(x|c, \theta^k) - v, \quad (1)$$

with P parts, threshold v, week hypothesis parameters θ^k, c denoting the object's center node's location and scale, and $F(I)$ the set of image features extracted by the Kadir & Brady feature detector. By constructing classifiers $f^*(I) = sign(f(I))$ based on part models at the basic level, we could group the objects into basic categories.

3. Discriminative attribute extraction

When we encounter a new object like a cup, it is not sensible to learn all the appearance details like most of the previous works did. It is enough to remember that it looks like a cup as well as its discriminative details. This can help to learn the visual appearance of new object types and speed the recognition process. We are inspired to mimic these behaviors in computer vision.

To describe the most discriminative visual details for each object class, we build a uniform generative attribute model combining several visual factors. Here we refer to the model framework in [5].

3.1. Multiple visual cues

Three visual cues are utilized, color, texture, and geometric information. Our analysis is all based on image segments.

We utilize a random sampling strategy to sample patches from all training images and generate two codebooks of patch texture as well as patch color respectively. For the color cue, we cluster patches in HSV space using K-Means. For the texture cue, the sampled patch images are convolved with MR8 filter bank to generate filter responses. Exemplar filter responses are chosen as textons via K-Means clustering.

Then, every pixel is soft-assigned to the patch color type. A segment is represented as a normalized histogram over the patch color types of the pixels it contains. By clustering the segment histograms from the training images we obtain a codebook A of segment color appearance. Each segment s is assigned to one appearance $a \in A$.

Given a segment of a training image, first convolve it with a filter bank and then label each filter response with a patch texture type which lies closest to it in filter response space. The histogram of textons is computed for each segment. By clustering the segment histograms from the training images we obtain a codebook T of segment texture.

For geometric properties, we only try to measure several simple properties of a segment in this paper: curvedness, compactness, elongation, area relative to the image, and symmetric properties.

3.2. The attribute model

We introduce an attribute model $M(\alpha, \beta, \tau, \{\lambda^j\})$ to describe discriminative visual details of class at lower levels. $\alpha(\tau)$ is the collection for all possible color (texture) appearance of segments with the attribute. $\alpha(\tau)$ can contain a single appearance, or all appearances in the codebook $A(T)$. The formal corresponds to specific color (texture) details, while the latter corresponds to generic patterns. For each geometric property $\lambda^j = (\phi_j, v_j)$, the model defines its distribution over the segments with discriminative details and whether the property is active or not ($v_j = 1$ or 0). β represents a background model. Figure 1 illustrates the conditional probability of segments. The Latent variable f is associated with each segment, indicating whether it is covered by the most discriminative details.

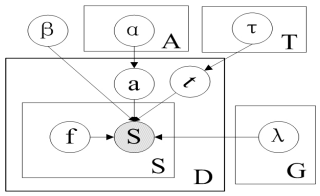

Figure 1. A graphical model for attributes

The conditional possibility of image segments could be computed as:

$$p(s|M; f, a, t) = \begin{cases} p(s_a|a)p(s_t|t)\prod_j p(s_g^j|\phi^j)^{v_j} & \text{if } f = 1 \\ \beta & \text{if } f = 0 \end{cases} \quad (2)$$

A segment s in I is defined by its color appearance s_a, texture s_t, and geometric characters $\{s_g^j\}$. If $s_a = a(s_t = t)$, then $p(s_a|a) = 1(p(s_t|t) = 1)$.

For each object class at a certain lower level, in the training stage, we learn the model parameters M that maximize the likelihood ratio in equation (3) with the training data consists of positive($\mathcal{I}_+$) and negative($\mathcal{I}_-$) images. Thus the most discriminative attribute for each specific category at this level is popped out. In the following equations, $N_s$indicates the number of pixels in s and we collect f for all segments of I into the vector $\mathbf{F}$.

$$\frac{p(\mathcal{I}_+|M)}{p(\mathcal{I}_-|M)} = \frac{\prod_{I_+^i \in \mathcal{I}_+} p(I_+^i|M)}{\prod_{I_-^i \in \mathcal{I}_-} p(I_-^i|M)}, \quad (3)$$

$$p(I|M) = \max_{a \in \alpha, t \in \tau} \max_{\mathbf{F}} p(I|M; \mathbf{F}, a, t), \quad (4)$$

$$p(I|M; \mathbf{F}, a, t) = \prod_{s \in I} p(s|M; f, a, t)^{N_s}. \quad (5)$$

The model parameters $(\alpha, \beta, \tau, \{\lambda^j\})$ are leant by a simple approximate optimization algorithm similar with [5]. Some learned models for representing discriminative details of classes are shown in figure 2. In (a), the most discriminative details of red and green apples are the specific color on the apple body. The three most frequent patch types of each specific color are given out. In (b), for tall cups and low cups, the geometric property of elongation is activated.

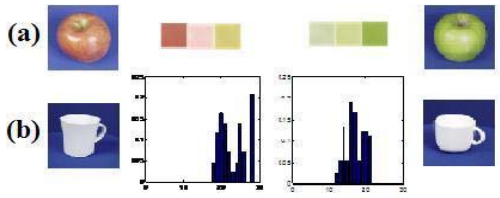

Figure 2. An illustration of learnt attribute models with some parameters

When given a test image and we are at one concept node A in the hierarchy of objects (lower than the basic level), the next subcategory labels would be given by:

$$C(I) = arg\max_{c \in children(A)} P(I|M_c). \quad (6)$$

4. Object category hierarchy

In this section we build a semantic hierarchy to integrate inference rules and all above classifiers for exploiting multi-level category information.

Figure 3 presents the graph of the hierarchy used in our experiments. Here to build this graph, we utilize WordNet tool [4] and refer to the process introduced in [7]. We also supplement tge graph with a few semantic relationships generated from the product catalog on some merchants' website. In this graph, the gray nodes represent the category labels at the basic level. For category labels interpretation, reasoning is possible using hypernymy(If it is an apple, it must be a fruit). Hypernymy relationships are also reflected in the figure 3.

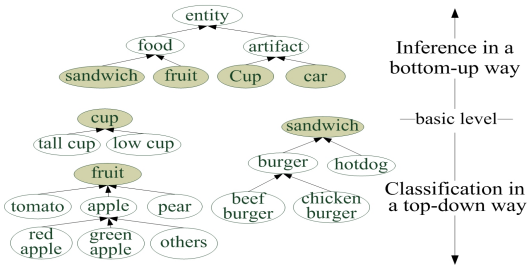

Figure 3. The hierarchy of categories in our dataset

To infer category information of higher levels, we use the prior information integrated in the semantic graph. Then for levels lower than the basic level, we train classifiers for each descending link. First we look for images supporting a given concept. Due to the hypernymy relationship, supposed B_i is a subcategory linking to A through hypernymy, we train a given $B_i|A$ classifier with

$$\mathcal{I}_+ = supp(B_i), \mathcal{I}_- = supp(A) - supp(B_i), \quad (7)$$

where $supp(X)$ is a set of images supporting the X concept. With training set grouped by equation (7), we could train a classifier for link $A \rightarrow B_i$.

So in the training stage, we train part-based class models at the basic level. Then at each lower level in the hierarchy, we model the most discriminative details of each class. Classifiers attached with the links in the graph are developed according to equation (1)and(6).

Given a test sample, first classified by part models, we get its basic concept label. Then we infer its upper category information by the prior semantic relationships ascendingly till reaching the entity node. For concept labels at lower levels, we descend to the linked concepts by the answer of corresponding classifier until reaching one leaf node. Finally we could get a set of category labels of this test sample at different levels.

5. Experiments

We designed a hierarchical data set first. As much as possible, classes were taken from standard benchmark datasets, with a few exceptions. We used images of the following categories in ETH80 dataset [6]: cup,

car, apple, pear and tomato. A set of sandwich images (burger and hotdog) are selected from the Caltech256 and Google Search to increase the complexity of the hierarchy. The above objects are further grouped or divided manually in a semantic way as shown in figure 3. Totally, the database contains 2350 images, of which some samples are shown in figure 4(a).

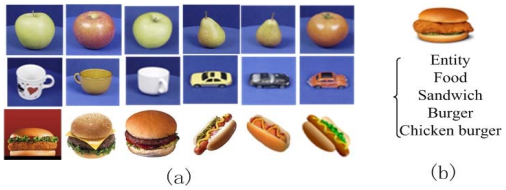

Figure 4. (a)Examples taken from the database. (b)An category interpretation result

We tested the proposed framework on the dataset. A sample with its interpreted multilevel category labels is shown in figure 4(b). For quantitive evaluation, we compute the precision rate of classification when a specific class label in the hierarchy is output. The results in table 1 show that our approach achieve comparative results with the state of art technique [6], while more semantic category information is exploited and the optimized decision cues are selected automatically instead of being set manually in [6].

Table 1. Multilevel categorization results

Multilevel categories		Our algorithm	Results in [6]
Higher levels	food	95.1%	-
	artifact	97.3%	-
Basic level	fruit	95.6%	-
	sandwich	92.0%	-
	cup	95.1%	96.1%
	car	99.5%	100%
Lower levels	tomato	97.6%	98.1%
	pear	99.0%	99.5%
	apple	90.1%	89.8%
	burger	83.3%	-
	hotdog	82.5%	-
	tall cup	81.7%	-
	low cup	93.5%	-
	chicken burger	70.0%	-
	beef burger	75.0%	-
	red apple	83.6%	-
	green apple	87.8%	-

6. Conclusions

In this paper, we have presented a framework for layered object categorizations. A semantic object hierarchy integrated with prior knowledge of relationships is built up first. Unlike the classical top-down and bottom-up classification method, following several human cognitive rules, we start to interpret an object from the basic level. Part-based models are used to classify categories at this level. Then different optimal cues are automatically selected for categorizations upwards and downwards simultaneously. Prior knowledge about inter-category relationships is utilized to infer objects' category labels ascendingly from the basic level. While the most discriminative visual details of each category are automatically selected for classification at the lower specific levels descendingly. A generative attribute model combining multiple visual cues to characterize visual details is utilized. We have tested our algorithm on a carefully designed hierarchical dataset and achieved promising results. In the future, we will try to include more complex local shape descriptors into the attribute model.

7. Acknowledgement

This research was partially supported by China Scholarship Council, NIH under Genes, Environment and Health Initiative (GEI) (Grant No.U01HL91736), National Basic Research Program of China (Grant No.2007CB311005), and National High-Tech Research and Development Plan of China (Grant No.2006AA01Z192). The work was performed in Carnegie Mellon University.

References

[1] A. Bar-Hillel and D. Weinshall. Subordinate class recognition using relational object models. *NIPS 2006*, 19:73–80.

[2] A. Bar-Hillel and D. Weinshall. Efficient learning of relational object class models. *IJCV*, 1-3(77):175–198, May 2008.

[3] A. Dhua and F. Cutzu. Hierarchical, generic to specific multi-class object recognition. *ICPR 2006*, 1:783–788.

[4] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass, 1998.

[5] V. Ferrari and A. Zisserman. Learning visual attributes. *NIPS2007*.

[6] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. *CVPR 2003*, 2:409–415.

[7] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. *CVPR 2007*, pages 1–7.

[8] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. *CVPR 2006*, pages 1447–1454.

[9] A. Zweig and D. Weinshal. Exploiting object hierarchy: Combining models from different category levels. *ICCV 2007*, pages 1–8.