Learning Consumer Photo Categories for Semantic Retrieval

Joo-Hwee Lim
Institute for Infocomm Research
21 Hcng Mui Keng Terrace, Singapore 119613
joohwee@i2r.a-star.edu.sg

Jesse S. Jin
University of New South Wales
Sydney 2052, Australia
jesse@cse.unsw.edu.au

Abstract

In this paper, wo develop a computational learning framework to build a hierarchy of 11 consumer photo categories for semantic retrieval. Two levels of visual semantics are learned for image content and image category statistically. We evaluate the average precisions at top retrieved photos on 2400 heterogeneous consumer photos with very good result.

1 Introduction

Research on image categorization has received more attention lately. In particular, the efforts to classify photos based on contents have been devoted to: indoor versus outdoor [Bradshaw, 2000], natural versus man-made [Bradshaw, 2000; Vailaya et al., 2001], and categories of natural scenes [Vailaya et al., 2001]. In general, the classifications were made based on low-level features such as color, edge directions etc and [Vailaya et al., 2001] presented the most comprehensive coverage of the problem with a hierarchy of 8 categories.

In this paper, we deal with a more comprehensive hierarchy of 11 categories (Fig. 1) for 2400 consumer photos. These photos (Fig. 2), including photos of bad quality, present a real spectrum of complexities when compared to the photos used in previous works [Bradshaw, 2000; Vailaya et al., 2001], which are mainly professional photos from the Corel stock photo library. Furthermore, only 20% of our test collection is used for training and we evaluate our approach in terms of average precisions (over 10 runs) of semantic retrieval at top retrieved photos which is important for practical usage. Our approach is unique that we compute *uniform semantic* features at both image content and image category levels using a computational learning framework instead of low-level features crafted for different categories.

At the image content level, salient image regions that exhibit semantic meanings are adopted as training examples to construct *semantic support regions* (SSR) that span a new indexing space. Local image regions of a photo is projected into this space as linear combinations of the SSR and further aggregated spatially to form image content signature for similarity matching. At the image category level, we learn image category models with small number of labeled photos to compute the relevance measure of photos in a winner-take-all approach.

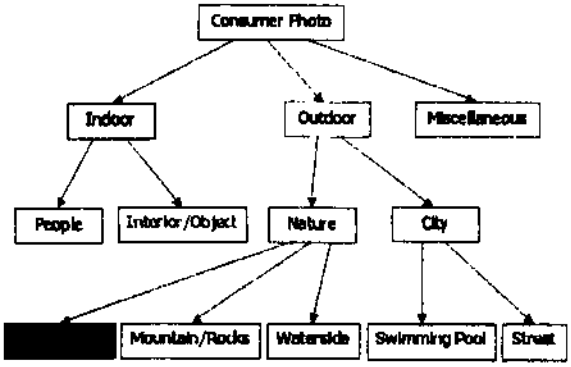

Figure 1: A hierarchy of consumer photo categories

1.1 Learning Semantic Support Regions

SSR are salient image patches that exhibit semantic meanings to us. A cropped face region, a typical grass patch, and a patch of swimming pool water etc can all be treated as their instances. To compute the SSR from training instances, we adopt support vector machines. We extract suitable features such as color and textures for a local image patch and denote this feature vector as *x*. A support vector classifier S_i devoted to a class *i* of SSR is treated as a function on x, $S_i(x)$. Then the posterior probability of class *i* can be computed as

$$P(S_i|x) = \frac{\exp^{S_i(x)}}{\sum_k \exp^{S_k(x)}}. \qquad (1)$$

For the experiments described in this paper, since we are dealing with heterogeneous consumer photos, we adopt color (means and standard deviations of each color channel) and texture features (means and standard deviations of Gabor coefficients) to characterize SSR.

To detect SSR with translation and scale invariance in an image, the image is scanned with windows of different scales. To reconcile the detection maps across different resolutions onto a common basis, we adopt the following principle: If the probability of the most probable class of a region *j* at resolution *k* is less than that of a larger region (at resolution *k*+ 1) that subsumes region *j,* then the classification probabilities of region *j* should be replaced by those of the larger region at resolution *k* -f 1. To aggregate the classification probabilities of the reconciled detection map for a spatial area *X* that comprises of *x* small equal regions with feature vectors $x_1, x_2, \cdots, x_n$, we compute the expected value of $\hat{P}(S_i|X)$ over x_3 as $\frac{1}{n}\sum_j P(S_i|x_j)$ since *P(XJ)* are equal for all small regions *j.* The similarity between two corresponding blocks X_3 in image *X* and *Yj* in image *Y* is computed as cosine between the probability vectors $\hat{P}(S_i|X_j)$ and $\hat{P}(S_i|Y_j)$. The overall similarity between *X* and *Y* is the average of the cosine values over all blocks.

1.2 Learning Semantic Image Categories

A photo category M_i is also learned using support vector machines. The input patterns to M_i are the indexes of the images $(Z = \hat{P}(S_i|X_j))$. The support vector learning computes the support images for the categories from a set of labeled photos. Given an unlabeled photo of index Z, the output of a category M_i is *S{Mj,Z).* With the winner-take-all approach, we compute the winner *k* as A: = *argmaxiS(Mi, Z).* Then the relevance measure of *Z* to category M, is defined as

$$R(M_i, Z) = S(M_k, Z) \quad if \;\; i = k; \;\; 0 \;\; otherwise \qquad (2)$$

2 Empirical Evaluation

From the 2400 photos, we define ground truth lists for the 11 categories. Their sizes are listed in Table 1 as breadth-first order of Fig. 1 and examples are shown in Fig. 2. We designed 2G classes of SSR: people (face, figure, crowd, skin), sky (clear, cloudy, blue), ground (floor, sand, grass), water (pool, pond, river), foliage (green, floral, branch), mountain (far, rocky), building (old, city, far), interior (wall, wooden, china, fabric, light). We cropped 554 image regions from 138 images and used 375 of them as training data for support vector machines to compute the SSR and the remaining 179 as test data to gauge generalization performance. Among all the kernels tried, a polynomial kernel with degree 2 and constant 1 gave the best result on precision and recall.

Figure 2: Two sample photos for each category (top-down, left-to-right): indr, outd, misc, inpp, inob, city, natr, pool, strt, wtsd, park, mtrk

We used 20% of the 2400 photos for training. We generated 10 different sets of positive training samples from the ground truth list for each category based on uniforrr random distribution. The negative training samples ofa given category are positive training samples from other categories that do not overlap with the category. The evaluation of retrieval precision is carried out hierarchically with respect to the category tree in Fig. 1. The test data for the category of a child node of a run is the ground truth list of its parent node minus the training samples used for learning the category of the child node- in the run. For example, to evaluate the retrieval performance of nature (natr) photos, the ground truth list of outdoor less the training sample for building the nature category is taken as the test data. The learning and retrieval of each category were performed 10 times and the average precisions (over 10 runs) of te>p retrieved images are given in Table 1.

Table 1: Average precisions at te)p numbers of photos

[Avg.Prec.	Size	Top 20	Top 30	Top 50
indr	994"	0.94	0.96	0.96
outd	1218	1.00	1.00	1.00
inpp	860	0.99	0.99	0.99
inob	134	0.84	0.75	0.56
natr	521	0.96	0.96	0.95
city	697	0.95	0.94	0.93
park	304	1.00	0.99	0.98
mtrk	67	0.41	0.27	0.16
wtsd	150	0.92	0.89	0.66
pool	52	0.47	0.32	0.21
strt	645	0.99	0.99	0.99

From Table 1, we observe that up to first 50 images, a user (on average) gets almost all relevant photos of the respective categories except less so for categories interior/object (inob) and waterside (wtsd), and even less sc for categories mountain/rocks (mtrk) and swimming poo! (pool). The reasons for poorer performance are 2-fold First these categories have much fewer positive training samples (i.e. 27,30,13,10). Next, they comprise images of varied contents (c.f.Fig. 2: objects plus interior, mountain and rocks, lakeside plus beach, pool wit! and without water as focus). We believe that with more training samples, their performance would be raised.

References

[Bradshaw, 2000] B. Bradshaw. Semantic based image retrieval: a probabilistic approach. In *Proc. of ACM Multimedia,* pp. 167-176, 2000.

[Vailaya et al., 2001] A. Vailaya et al. Bayesian framework for hierarchical semantic classification of vacation images. *IEEE Trans, on Image Processing* 10(1): 117-130, 2001.