Sampling and Ontologically Pooling Web Images for Visual Concept Learning

Shiai Zhu, Chong-Wah Ngo, and Yu-Gang Jiang

***Abstract*—Sufficient training examples are essential for effective learning of semantic visual concepts. In practice, however, acquiring noise-free training examples has always been expensive. Recently the rapid popularization of social media websites, such as Flickr, has made it possible to collect training exemplars without human assistance. This paper proposes a novel and efficient approach to collect training samples from the noisily tagged Web images for visual concept learning, where we try to maximize two important criteria, *relevancy* and *coverage*, of the automatically generated training sets. For the former, a simple method named semantic field is introduced to handle the imprecise and incomplete image tags. Specifically, the relevancy of an image to a target concept is predicted by collectively analyzing the associated tag list of the image using two knowledge sources: WordNet corpus and statistics from Flickr.com. To boost the coverage or diversity of the training sets, we further propose an ontology-based hierarchical pooling method to collect samples not only based on the target concept alone, but also from ontologically neighboring concepts. Extensive experiments on three different datasets (NUS-WIDE, PASCAL VOC, and ImageNet) demonstrate the effectiveness of our proposed approach, producing competitive performance even when comparing with concept classifiers learned using expert-labeled training examples.**

***Index Terms*—Training set construction, visual concept learning, web images.**

I. Introduction

VISUAL concept learning has recently received great attention in multimedia and computer vision. It is fundamentally a classification task that determines whether a concept (e.g., an object or scene class) is present in images or video clips. A critical step for learning effective concept classifiers is the acquisition of a sufficiently large number of training examples. Manual collection of these training data is often expensive and infeasible when the number of concepts is very large. Existing human annotation efforts like TRECVID [1], PASCAL VOC [2], and ImageNet [3] are lacking in either label accuracy or diversity/coverage, limiting the needed progress in this vibrant field.

Manuscript received June 22, 2011; revised October 28, 2011 and January 19, 2012; accepted February 24, 2012. Date of publication March 08, 2012; date of current version July 13, 2012. This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under Grant CityU 119709. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marcel Worring.

S. Zhu and C.-W. Ngo are with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong (e-mail: shiaizhu2@student.cityu.edu.hk; cwngo@cs.cityu.edu.hk).

Y.-G. Jiang is with the School of Computer Science, Fudan University, Shanghai 201203, China (e-mail: ygj@fudan.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2012.2190387

Fig. 1. Example Web images with noisy yet informative tags. This paper proposes novel and efficient techniques to automatically sample the noisily tagged Web images for learning semantic visual concepts. (a) Subjective tags. (b) Ambiguous tags. (c) Incomplete tags. (d) Content and context related tags.

Recently, with the proliferation of social media websites, more and more images with user tags are freely available on the Web. Flickr.com, for example, receives thousands of new uploads per minute. The associated user-generated tags of these images convey useful information on their semantic content. This gives light to explore the Web, an extremely rich data source, for constructing concept training sets. Since this process is fully automatic and can be repeated easily, we can always generate new training sets with *up-to-date* image content, in contrast to those manually labeled collections like ImageNet where the content of some concepts may become outdated after a few years (e.g., the appearance of "computer monitor" has changed a lot over the years).

The tags of Web images, nevertheless, are usually imprecise and incomplete. A recent study by Kennedy *et al.* [4] showed that only 50% of the tags are visually relevant to image contents. Fig. 1 shows some example images from Flickr, based on which we summarize the challenges of sampling Web images for concept learning as follows. First, user-tags tend to be subjective and highly personalized. Tag providers are mostly the owners of the images and therefore may supply tags that are not necessarily content-related. For example, in Fig. 1(a), "child" is visually very relevant but is not in the tag list. Instead, other tags such as "barossa" and "valley," indicating the place or location where the photo was taken, are used. Using this example for training concepts such as "valley" could result in worse performance. Second, user tags can be ambiguous. Some tags may convey multiple senses or meanings. For instance, in Fig. 1(b), while "bear" is commonly referred to as a kind of animal, it also has another meaning: "a surly, uncouth, burly, or shambling person." Third, lists are often incomplete. Taking Fig. 1(c) as an example, the tag "beach" is missing, while some other relevant

tags such as "sea," "sky," and "sand" are tagged. In this case, sampling training images merely based on keyword matching can lead to a low recall of positive samples. Last, some tags are more content specific and representative than others. For example, the image in Fig. 1(d) can be a representative example for learning concept "dog" since many contextual tags like "animal," "pet" and "puppy" give a strong clue that "dog" is the major highlight of this image.

This paper proposes a novel and efficient solution to sample Web images for visual concept learning. Particularly, we consider the following two important requirements when constructing the concept training sets: 1) the noisily tagged Web images should be filtered to **maximize** their **relevancy** to a target concept, and 2) the final training set should have **good coverage and diversity**. To deal with the challenges described earlier, we introduce a method called semantic field (SF) to determine the relevancy of an image to a target concept, so that the noise in the constructed training set can be suppressed. Different from direct text matching between image tags and concept names, in SF we consider tags of an image collectively to predict image semantics. This way we are able to handle most issues raised by the examples in Fig. 1, which will be elaborated in Section III-A. Having a noise-free training set is desirable but not sufficient for training a good concept classifier. As stated in requirement 2, we also need a training set diverse enough to cover most visual aspects of the target concept. For instance, images of both "house" and "church" are needed for training a comprehensive and generalizable classifier for concept "building." To this end, we propose semantic pooling (SP), a simple technique for boosting the coverage of concept training sets by propagating positive examples across semantically related concepts using an ontology structure. With the ontology-based SP, training set of a non-leaf concept (e.g., "building") can be enriched by including representative examples from its child nodes (e.g., "church").

Our main contribution in this paper is a systematic approach to sampling Web images for concept learning. Both SF and SP are easy to implement and scalable to large scale applications. Thorough evaluations are conducted over several popular datasets to validate the effectiveness of our approach.

The remainder of this paper is organized as follows. In Section II, we review existing works on concept learning and training set construction. Section III elaborates the proposed SF and SP methods for collecting training images from the Web. Section IV describes the experimental setup and discusses evaluation results. Finally, Section V concludes this paper.

II. Related Works

A. Visual Concept Learning

Numerous research efforts have been devoted to visual concept learning in images and videos. This subsection briefly discusses feature representation and model learning techniques.

Recent research has shown that bag-of-visual-words representation computed based on local image features (e.g., SIFT [5]) has converged to a mature solution for concept learning [6], [7]. Many top-performing systems at various benchmark evaluations purely relied on this feature representation [2], [8], [9]. For learning techniques, SVMs are the dominant choice provided that a sufficient amount of training samples is available. When the number of training samples is small, semi-supervised learning which explores freely available unlabeled data can be applied. For example, Yan and Naphade [10] proposed a multi-view semi-supervised cross-feature learning approach. Initially one classifier from each view is learned by expert-labeled training data. The model is further boosted by augmenting the training set of one view with selected unlabeled testing data on which the other views have high-confidence prediction. However, Tian *et al.* [11] have pointed out that detection performance of semi-supervised learning methods will degrade when labeled and unlabeled data are from different distributions. In fact, the change of data distribution can affect not only semi-supervised learning, but also supervised learning [12].

In the following we discuss prior works on constructing training sets, which are highly related to the main focus of this paper. For features and model learning, we simply adopt the cutting-edge solutions of bag-of-visual-words feature and SVMs classifier.

B. Concept Training Set Construction

Collecting sufficiently large amount of training data always plays an important role in learning robust visual concept classifiers. To date, most of the released datasets are constructed by expensive manual labeling, despite the fact that automatic sampling of web images for learning is feasible with the massive growth of user-tagged images on social media websites. In this section, we divide the review of prior works on training set construction into two categories: manual annotation and automatic collection from Web.

Many efforts have been devoted to manually annotating images for visual learning. Some small yet very popular datatsets include Caltech-256 [13], MSRC [14], LabelMe[1] [15], MIR-Flickr [16], etc. The average number of labeled samples per concept is quite small in these datasets. Recent developments like the PASCAL VOC 2010 [2] provided more examples per class, but usually only have a very small number of categories due to the limitation of annotation human-hours (20 concepts in VOC 2010). The NUS-WIDE dataset [17] includes about 270 000 Flickr images with ground-truth annotation of 81 concepts, which have a better coverage of the visual world surrounding us but are still far from adequate. ImageNet [3] dataset, in contrast, has more than 12 million images manually labeled to 17 000 concepts, and is still growing. On average, there are about a few hundreds of examples per concept, which are still not always enough for learning good visual classifiers. In the video domain, the LSCOM [18] has annotated more than 400

[1]LabelMe is a Web-based annotation tool developed by MIT CSAI Lab to collect object-level image annotations from general Web users.

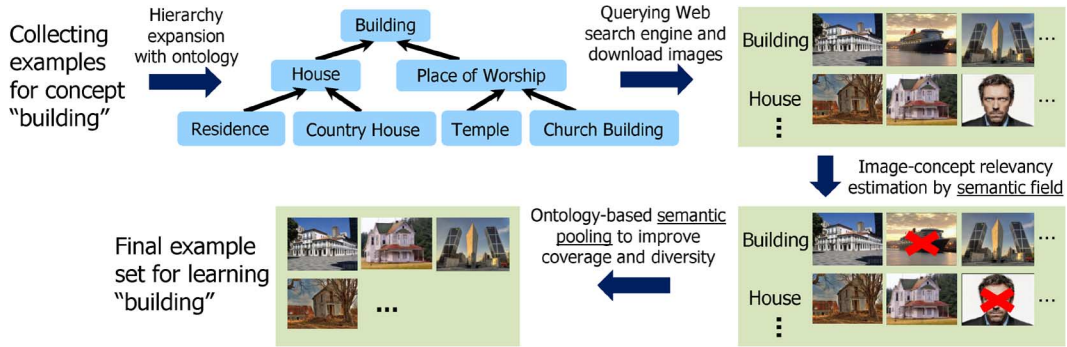


Fig. 2. Overall framework of the proposed approach. Given a concept, WordNet ontology is first adopted to produce a hierarchy with the target concept as root node. Each node (concept) in the hierarchy is then used to query Web search engines and download top-ranked images. After that, the downloaded images of each node are filtered using a method called SF. Finally, example set of the target concept is constructed by pooling examples from all the nodes in the hierarchy.

concepts, but is restricted to broadcast news videos. In addition, through collaborative annotation, there are also 346 concepts being labeled on a set of Web videos provided by NIST TRECVID[2] 2011 [1].

Automatically collecting training images by sampling social media becomes a very natural and plausible solution due to the popularity of social media websites [19]–[25], [33]. Among these, one common and straightforward way is to directly use the top-ranked images from Web search for classifier learning [19], [20]. For instance, the MIT 80 million TinyImage dataset [26] was constructed by searching around 53 000 nouns defined by WordNet in various image search engines and crawling the top-ranked ones. Since the tiny images were directly downloaded from Web search without any filtering effort, the high level of label noise (around 20% are clean images) largely limits its utility for general algorithm development. In [27], the problem is dealt with by using probabilistic latent semantic analysis (pLSA). They assumed that the most relevant images are in several large clusters (topics), which will be used for reranking the search results. In [22] and [28], a two-step learning approach, combining context and visual information, was proposed to harvest image databases from Web. First, surrounding context of images from initial Web search is used to train a Bayes posterior estimator which is used for reranking the search results. After that the top ranked images are sampled to train a SVM classifier to further refine the image ranking. In [23], starting from a classifier model learned from a collection of expert-labeled examples, iterative updates of the model are performed by training with new web images that are predicted as positive by the previous older model. At the end a larger training set will be automatically constructed. The newly added images by this method, however, may be visually homogeneous and lack visual diversity. In [29], instead of ranking individual images as in [33] and [21]–[23], image clustering of search result is firstly executed and then random walk is applied to rank the image clusters. Simple heuristics such as larger clusters are more relevant to the target concept and noisy samples should be outliers from clustering are adopted for training set construction. In [30] and [31], a filtering approach is proposed to sample Web videos for concept learning. Based on a probabilistic framework, the relevancy of a video frame to a target concept was learnt in order to exclude irrelevant frames from concept training. To improve the performance of frame sampling, in [32], the framework was further extended by using active learning.

Because tags in social media could be noisy, semi-supervised learning, which starts by learning classifiers from a small set of manual labels and subsequently updates the classifiers by sampling web images, is adopted by [21] and [33]. Many other works focused purely on refining the user-supplied image tags [34]–[36], which all used image visual similarities with simple models. In addition to sampling positive examples, the problem of negative sample selection is also studied by [24] and [25]. In [24], random sampling was simply applied. In [25], a bootstrap learning approach was proposed, by iteratively selecting the most misclassified images of a concept to update the classifier.

Different from tag refinement by learning tag relevance from neighborhood visual similarities as in [34]–[36], our work in this paper targets at sampling Web images for concept learning by collectively analyzing the whole tag list. In addition, different from [19]–[23], [29], and [33], we not only focus on sampling relevant images, but also aim to construct a comprehensive training set that can cover different visual aspects of a target concept using ontological pooling. Different from a recent work by Fergus *et al.* [37] where an ontology was adopted to expand query sets for label propagation, we use ontology to pool comprehensive sets of training samples for concept learning. In addition, our training set construction approach does not involve any model learning process, and therefore is highly efficient, satisfying the critical speed requirement of many large-scale applications. This work extends a previous conference paper [38] with an algorithm for SP and an extensive set of newly added experiments. Additionally, we also conduct a cross-dataset evaluation to test the generalization capability of models learnt using various kinds of training labels.

III. Sampling and Ontologically Pooling Web Images

This section introduces our proposed methods for concept training set construction. In particular, we focus on collecting comprehensive sets of positive samples with good relevancy and coverage. For negative sample selection, we simply adopt the random sampling method of [24].

[2]TRECVID is an annual video retrieval benchmark organized by the U.S. National Institute of Standards and Technology.

Fig. 2 shows the overall framework of our approach. Given a target concept, we first use the WordNet ontology to produce a concept hierarchy, with the target concept as root node. Each concept in the hierarchy is used to query Web search engines and download top-ranked images. Noises in Web search results are then suppressed using a method called SF. Finally, by pooling the pseudo examples of each node in the hierarchy, final example set for the target concept is formed. Key components in our approach include SF image sampling and hierarchical example pooling. We elaborate each of them below.

A. Modeling SF

Current image search engines largely rely on the associated texts (e.g., tags) of the images, and therefore often return results at low precision due to the noisy tags. While user tags are imprecise individually, inferring image semantics from tags is still feasible by collectively analyzing all the tags of an image together, with a reasonable assumption that generally the number of content-related tags is larger than that of the noisy tags. SF is proposed under this scenario to predict the relevancy of an image to a concept. Originally the concept of SF was used to capture the semantics of a set of words in text domain [39]. The basic idea is that the meaning of a word partially depends on its surrounding words. Under our application, a tag list can be treated as a SF, where the dominant semantics can be inferred by linguistic analysis of the tags.

Denote C_x as the target concept, and $SF = \langle T_1, T_2, \ldots, T_n \rangle$ as the tag list of an image I with n tags. By Bayesian theorem, the probability of C_x appearing in I can be formulated as

$$P(C_x|SF) = \frac{P(SF, C_x)}{P(SF)}. \tag{1}$$

The computation of (1), however, is not always stable since the probability of the entire tag list $P(SF)$ is usually extremely small. We thus adopt simple average fusion to approximate $P(SF, C_x)$ by $P(SF) \times (\sum_i P(T_i|C_x)/n)$, which combines the probabilities of observing SF as a whole and averaging the probability of each tag being relevant to concept C_x. Here average fusion of probabilities is used since the tags of an image are usually not independent (highly correlated sometimes), which makes it infeasible to follow a strict probabilistic framework for computing $P(SF, C_x)$. With this approximation, $P(SF)$ can be eliminated and (1) can be rewritten as

$$P(C_x|SF) = \frac{\sum_{i=1}^{n} P(T_i|C_x)}{n} \tag{2}$$

where $P(T_i|C_x)$ donates the likelihood of observing a tag T_i given the concept C_x. Since $P(T_i|C_x)$ in (2) can be further rewritten as $P(T_i, C_x)/P(C_x)$, and $P(C_x)$ does not affect image sampling for C_x, the only critical unknown term for computing $P(C_x|SF)$ is the joint probability $P(T_i, C_x)$.

To estimate $P(T_i, C_x)$, we consider two different knowledge sources: WordNet ontology and Flickr.com. For WordNet, we adopt WUP [40] which uses path length information in WordNet hierarchy to infer word relatedness, defined as

$$WUP(T_i, T_{C_x}) = \frac{2D(S_{T_i,T_{C_x}})}{L(T_i, T_{C_x}) + 2D(S_{T_i,T_{C_x}})} \tag{3}$$

where T_{C_x} denotes the name of concept C_x and $S_{T_i,T_{C_x}}$ is the lowest common ancestor of T_i and T_{C_x} in WordNet. Function D returns the depth of a concept, while function L computes the minimum path length by traversing from T_i to T_{C_x}. WUP does not exactly consider joint probability, but reflects the relationship of the two words from semantic point of view. In the implementation, we only consider tags which are found in WordNet for WUP computation. Tags which cannot be found in WordNet include abbreviation, number, wrongly-spelled, slang (e.g., beemer) and linked (e.g., johnaryanphotography) words. Among all the images crawled from Flickr in our experiments, about 37% of the tags cannot be found in WordNet.

In addition to WUP, we estimate the relationship of T_i and T_{C_x} based on the co-occurrence of both words. We adopt Flickr Context Similarity (FCS) [41] which estimates the co-occurrence of tags based on statistics derived from tags associated with all images in Flickr. This offers the advantage that the co-occurrence of words could also reflect visual relatedness since tags are given with images as the target subjects. FCS is defined as

$$FCS(T_i, T_{C_x}) = e^{-NGD(T_i,T_{C_x})/\rho} \tag{4}$$

where

$$NGD(T_i, T_{C_x}) = \frac{\max\{\log h(T_i), \log h(T_{C_x})\} - \log h(T_i, T_{C_x})}{\log N - \min\{\log h(T_i), \log h(T_{C_x})\}}. \tag{5}$$

Here NGD stands for Normalized Google Distance [42], $h(T_i)$ is the number of Flickr images associated with tag T_i, $h(T_i, T_{C_x})$ is the number of images associated with both T_i and T_{C_x}, and N is the number of images indexed by Flickr. The function $h()$ is computed by querying Flickr API. Based on [41], the parameter ρ in (4) is empirically set to 0.25.

Finally, with WUP and FCS, $P(T_i, C_x)$ can be estimated by

$$P(T_i, C_x) = FCS(T_i, T_{C_x}) \times WUP(T_i, T_{C_x}) \tag{6}$$

which aims to boost the rank of a tag if receiving high scores from both WUP and FCS. Plugging (6) back into (2), $P(C_x|SF)$ can be computed for each image under consideration, with which images from initial Web search are reranked and top ones will be selected for training set construction.

B. Ontology-Based SP

In WordNet ontology hierarchy, child nodes are semantic subsets of parent nodes. Take the concept "building" as an example, using the hyponymy relationship in WordNet, nodes under "building" are organized in a sub-tree structure of six layers and 268 child nodes. Intuitively, the coverage and diversity of training examples for "building" can be greatly enhanced, by pooling examples of the child nodes. With this intuition, we now adopt the WordNet ontology to propagate positive examples sampled by SF for concept training set construction.

Fig. 3 shows the top three levels of the ontology for concept "building". SP is performed in bottom-up manner so that a concept can receive examples from its child nodes. Specifically, taking a tree with only two layers as an example, positive samples from the child nodes are propagated in proportion to the root node. The proportion is decided based on the popularity of

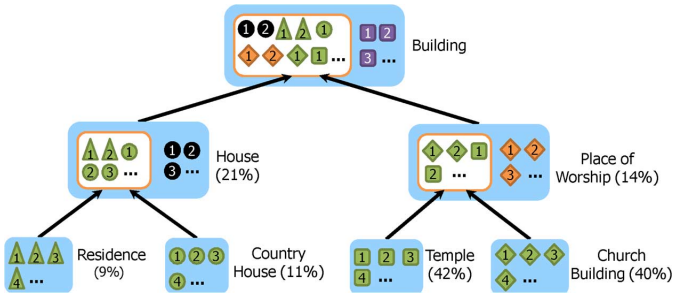


Fig. 3. Toy example of SP of training examples for concept "building." Positive samples of the child nodes are hierarchically pooled in bottom-up manner. The percentage in the parentheses indicates the proportion of examples to be propagated to the parent node, computed by (7). The numbered small color boxes (in online version) with various shapes represent images originally from different nodes, with rankings (indicated by the numbers) computed by (2).

the child node, which is measured based on the total number of images returned by Flickr API. Formally, the percentage of training samples to be propagated from a child node C_i to its parent node can be computed by

$$P_{C_i} = \frac{f_{C_i}}{\sum_{C_j \in L_i} f_{C_j}} \tag{7}$$

where f_{C_i} is the number of Flickr images tagged with C_i, and L_i represents a set of concept nodes on the same layer with C_i in the tree.

The order of selecting samples for pooling is based on the rank list evaluated by (2). In other words, the first image being picked up from a child node is always its top-ranked image estimated by SF. Images propagated into a parent node are rescored by the following equation:

$$S(I_k) = \frac{(N-k)}{N} \tag{8}$$

where I_k is the image ranked at the kth position among training examples of the child node C_i, and N is the number of images being propagated from C_i. After the bottom-up propagation process, samples arriving at the root node C^* are then aggregated with the original samples T_{C^*} in C^* as follows:

$$\widetilde{T}_{C^*} \leftarrow T_{C^*} \cup T_{C_1} \cup T_{C_2} \cup \cdots \tag{9}$$

where T_{C_i} denotes the set of positive examples propagated from the child node C_i, and $\widetilde{T}_{C^*}$ is the final set of examples for learning concept C^*. To facilitate successive propagations into nodes on the upper layers, the examples in $\widetilde{T}_{C^*}$ is then ranked by (8), which guarantees that all the top-ranked examples of the child nodes will be distributed evenly in the rank list of $\widetilde{T}_{C^*}$ and the orders of images from the same child node can be retained. For a tree with more than two levels, similar procedure is carried out recursively from leaf nodes to the root concept. A toy example illustrating the procedure of SP for concept "building" is given in Fig. 3.

In (9), the number of training examples required for classifier learning is difficult to predict in practice, and so is the number of examples that should be pooled from child nodes for learning. A reasonable strategy is that the examples contributed by child nodes should not be more than that of the target concept. Therefore, we empirically choose the setting that both target and child nodes contribute equally to the number of training examples in the experiments.

C. Complexity and Implementation Details

With the formulations of SF and pooling, the framework shown in Fig. 2 is easy to implement and scales linearly with the number of concepts and images. The complexity is mainly governed by the computation of SF which evaluates the relationship of each tag with the target concept. We implement SF by offline learning a dictionary for each concept. First, a large pool of image tags is crawled. Second, the dictionary of a concept is built by keeping the top-k tags ranked using (6). Specifically, each entry of the dictionary is composed of a tag T_i and the relevancy $P(T_i, C_x)$. During sampling, given an image and its tag list, (2) can be efficiently computed by dictionary look-up and score averaging. To guarantee fast look-up, we only keep the top-200 tags in a dictionary, and set $P(T_i, C_x) = 0$ for the remaining tags. Because the ranking scores of tags usually drop exponentially, keeping an excessive number of tags in a dictionary will not improve sampling performance in practice.

IV. Experiments

A. Data, Classification, and Evaluation Criteria

The major dataset used to evaluate our proposed approach is the NUS-WIDE Web image collection [17]. NUS-WIDE is composed of 269 648 Flickr images, each of which has 23 tags on average. Its training set contains 161 789 images and the test set has 107 859 images. Images of NUS-WIDE have been fully annotated with 81 concepts that can be roughly divided into six categories: people, objects, scene/location, event/activities, program, and graphics.

To test the generalizability of concept classifiers trained by our automatic example sampling approach, we also experiment with two additional datasets: PASCAL VOC 2010 [2] and a subset of ImageNet [3]. The 2010 edition of PASCAL VOC Challenge released a dataset of 19 740 Web images with labels of 20 concepts. The dataset is divided evenly into a training set and a test set. ImageNet has more than 10 000 concepts, each with a set of manually annotated examples. In this work, for the ease of cross-dataset evaluation, we use 30 concepts that overlap with the 81 classes labeled on NUS-WIDE. In total, ImageNet30 has 42 810 images which are evenly partitioned for training and testing.

For SF modeling, a dictionary of tags is learnt for each concept. The score of each tag, i.e., $P(T_i, C_x)$ in (6), is computed using Flickr.com (not NUS-WIDE) and WordNet. Visual concept classifiers are learned based on the settings of VIREO-374 [7], a popular baseline of concept learning. Three SVM classifiers are trained separately using bag-of-visual-words (BoW) feature, grid-based color moment and wavelet texture. Details of feature extraction can be found in [7]. For testing, average later fusion is used to combine probability predictions of the three SVM classifiers. Test images can then be ranked according to

TABLE I
RELEVANCE OF TRAINING SAMPLES COLLECTED BY VARIOUS METHODS, MEASURED BY BOTH RECALL AND MAP. THE 81 NUS-WIDE CONCEPTS ARE GROUPED INTO SIX CATEGORIES

Categories	Recall				MAP			
	SF	SF-FCS	SF-WUP	KW	SF	SF-FCS	SF-WUP	KW
people	**0.543**	0.529	0.474	0.447	**0.695**	0.688	0.666	0.536
objects	**0.548**	0.539	0.496	0.480	**0.680**	0.677	0.640	0.574
scene/location	**0.427**	0.419	0.375	0.360	**0.580**	0.572	0.565	0.518
event/activities	**0.436**	0.418	0.400	0.300	**0.510**	0.485	0.435	0.365
program	**0.670**	0.660	0.661	0.178	**0.705**	0.695	0.684	0.278
graphics	**0.175**	0.150	0.100	0.000	**0.140**	0.136	0.132	0.000
all concepts	**0.483**	0.473	0.432	0.397	**0.615**	0.607	0.582	0.515

the fused SVM predictions that indicate the confidence of detecting a concept in the images.

As the standard evaluation criterion of concept learning, average precision (AP) is adopted to assess the performance of concept classifiers. Mean AP (MAP) is used to aggregate performance over multiple concepts. To assess the relevance of the automatically sampled training images, we use Recall to evaluate the fraction of ground-truth positive examples being sampled, and AP to measure the ranking performance.

B. Experiment 1: Relevance of Training Examples

In this experiment, we evaluate the SF method for sampling positive examples from noisy Web search. This process is simulated by treating the training set of NUS-WIDE as the candidate image pool.

We compare SF with a baseline sampling method using simple keyword matching (KW) and also the expert-labeled examples from NUS-WIDE. The completely manual labeled training sets can be viewed as the upper limit of automatic sampling. For the computation of SF, in addition to using both WUP and FCS as shown in (6), we also investigate the sampling performance using FCS or WUP alone (named as SF-FCS and SF-WUP). To make sure the comparison is fair, we assume the number of positive examples is known, and when sampling the top-k images using SF and KW, k is set equal to the number of true positives according to the expert labels. For negative examples, throughout the experiments we randomly sample a fixed number of 5000 images, where the determined positive examples are excluded.

First, we measure the relevance of training examples by Recall and MAP. Table I summarizes the results. Both SF-FCS and SF-WUP perform better than KW. Further combination of the two knowledge sources (SF) improves the relevance of training examples moderately, with Recall and MAP equal to 0.483 and 0.615 respectively. These results clearly validate the importance of collectively analyzing the tags of an image. SF consistently gives the best performance across all the six categories of concepts. Note that SF exhibits the largest margin of improvement compared to the baseline for the category "program", and the lowest Recall and MAP for "graphics" compared to other five categories. The reason is due to the fact that both categories contain only one concept, respectively, the concepts "sports" and "map." For "sports," SF is effective in sampling images that are tagged with sport related words like "athletics" and "racing." In contrast, KW which uses simple keyword matching can only include images tagged with "sports" or "sport." The concept "map" has multiple senses and all the methods perform poorly. While not fully satisfactory, SF can still recall 17.5% of true positives.

TABLE II
MOST RELEVANT TAGS TO CONCEPT "AIRPORT," DETERMINED BY FCS, WUP, AND THEIR COMBINATION

	vocabulary				
FCS × WUP	airport	aircraft	runway	terminal	international
	landing	plane	airfield	airline	airplane
FCS	airport	airplane	aircraft	plane	airline
	aviation	international	flight	airbus	terminal
WUP	airport	airfield	zoo	station	transportation
	museum	gym	base	terminal	archive

Comparing SF-FCS with SF-WUP, SF-FCS is significantly better. This confirms the fact that statistics from Flickr can reflect visual co-occurrence of words more accurately. Table II lists the top 10 highly related tags to concept "airport," predicted by FCS, WUP, and their combination $(\mathrm{FCS} \times \mathrm{WUP})$. As shown in the table, FCS is able to select tags that visually co-occur with "airport," such as "airplane," "aircraft," and "airline." On the other hand, WUP tends to choose semantically related tags such as "airfield," "station," and "transportation," though not without misleading tags. Examples include "zoo" which is a sibling concept of "airfield" and shares a common ancestor "installation" with "airport." As a result, "zoo" is also ranked high in the vocabulary. By combining WUP and FCS as SF does, the top ranked tags are more diverse and comprehensive.

Next, let us measure concept learning performance based on different training sets. Table III shows the results. The classification performance of SF $(\mathrm{MAP} = 0.166)$ shows a closer MAP to that of using noise-free expert labeled training data $(\mathrm{MAP} = 0.222)$, and improves significantly over KW $(\mathrm{MAP} = 0.124)$, with a relative performance gain of 33.8%. The improvement is consistently observed in all the six categories of concepts. This result again confirms the effectiveness of collective tag analysis, since KW can be easily affected by individual noisy tags.

C. Experiment 2: Combining SF and SP

Instead of sampling positive examples from NUS-WIDE, in this experiment example sampling is conducted directly on images crawled from Flickr. Among the 81 concepts in NUS-WIDE, we consider 38 concepts who have at least two layers of child nodes according to the WordNet hierarchy. Among the 38 concept trees, the average depth is four and the average number of child nodes is 54. For instance, concept "building" has 127 nodes distributed over a tree of six layers, and "sports" has a seven-layer tree of 101 nodes. For each node of the concept

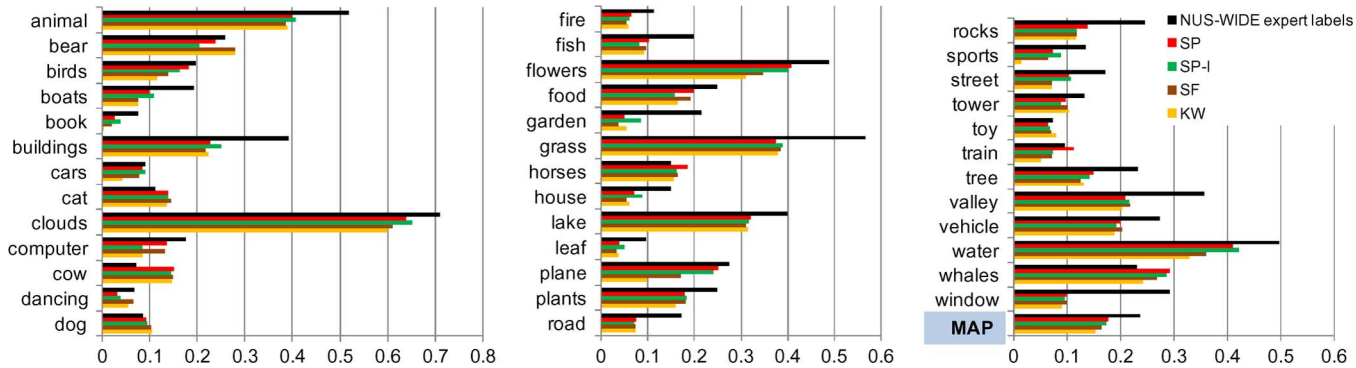


Fig. 4. Perconcept AP on NUS-WIDE test set, using models trained with various Web image sampling and pooling methods. See texts for detailed explanations.

TABLE III
CONCEPT LEARNING PERFORMANCE USING DIFFERENT EXAMPLE SAMPLING METHODS

Categories	MAP		
	Expert Labels	SF	KW
people	0.209	**0.143**	0.096
objects	0.196	**0.163**	0.129
scene/location	0.284	**0.210**	0.156
event/activities	0.110	**0.050**	0.021
program	0.135	**0.063**	0.024
graphics	0.180	**0.028**	0.000
all concepts	0.222	**0.166**	0.124

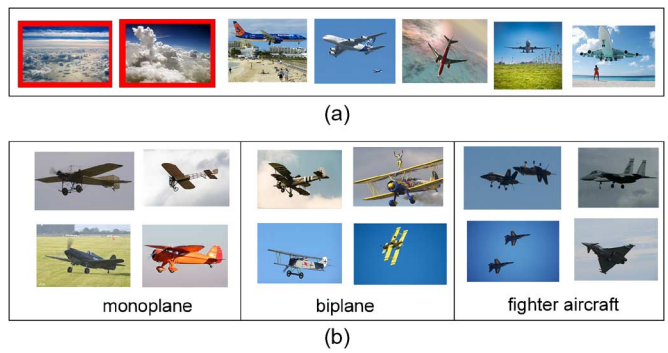


Fig. 5. Automatically sampled training examples for "plane" and its child concepts. False positive examples are marked with red boxes (in online version). (a) Training samples for "plane." (b) Training samples for child nodes.

trees, we download a large set of images by using the concept name to query Flickr API.

We compare the following approaches: SF, SP, and SP with incremental learning (SP-I). SF samples images from Flickr for each concept. SP, in addition to the sampled images of a concept by SF, further pools the images collected by child nodes as training examples, as described in Section III-B. SF and SP then learn the classifers based on the collected images respectively. Different from SP, SP-I updates the original classifiers learnt by SF using the new positive examples pooled from child nodes. We adopt adaptive SVM (A-SVM) [12] for incremental learning. A-SVM learns a "delta function" $\Delta f(x)$ based on the new examples, and adapts the original SVM model $f^a(x)$ as follows:

$$f(x) = f^a(x) + \Delta f(x) = f^a(x) + W^T \phi(x) \qquad (10)$$

where W^T are the parameters to be leant from new samples. A-SVM basically seeks for additional support vectors learnt from newly arrived data to adjust the original decision boundary of a classifier. It optimizes the trade-off that new decision boundary should be close to the original one, and meanwhile, the new samples are correctly classified.

Different from Experiment 1 where we assume the number of positive images is known for each concept, under the more realistic setting of crawling images directly from the Web, this number can never be accurately estimated. Therefore we adopt a very common strategy by sampling a fixed number of positive images for all the concepts. All the methods being compared select 2000 examples per concept. For SP/SP-I, the first 1000 examples are from the root node (target concept), and the remaining 1000 examples are pooled from the child nodes.

Fig. 4 shows the concept learning results on NUS-WIDE testing set. SF, with a MAP of 0.165, exhibits better performance than KW ($\mathrm{MAP} = 0.153$). This is consistent with the observations in Section IV-B. With SP, the MAP is further boosted to 0.176. Among the 38 tested concepts, 27 of them benefit from pooling examples of child nodes, and only five concepts suffer from performance degradation after pooling (by around 5%). The performance degradation for these concepts is mostly due to mismatch of concept definition between WordNet and image corpus. For example, the ground-truth labels in NUS-WIDE view "panda" as a kind of "bear," while in WordNet, "panda" is not regarded as a child node of "bear." On the other hand, "beach ball" and "tree house" are two child nodes of "toy" in WordNet, but are rarely labeled as "toy" in Web image corpus.

Overall, the coverage and diversity of training examples for most concepts can be enhanced by SP. Fig. 5 shows an example for concept "plane." In Fig. 5(a), the sampled examples for root node "plane" are mostly from close-up viewpoint and mixed with a few false positives. The samples from child nodes, as shown in Fig. 5(b), offer a more diverse view in terms of visual and semantic aspects. In addition, as the child node concepts are more specific, the chance of sampling false positives is usually lower than that for the parent concepts. Fig. 6 further gives top-60 plane classification results using SF and SP labels respectively. We see that SF [Fig. 6(a)] can find some close-view planes, but some sky or cloud images are also mistakenly included. In contrast, top results by SP labels [Fig. 6(b)] are more accurate and also visually diverse, though not without

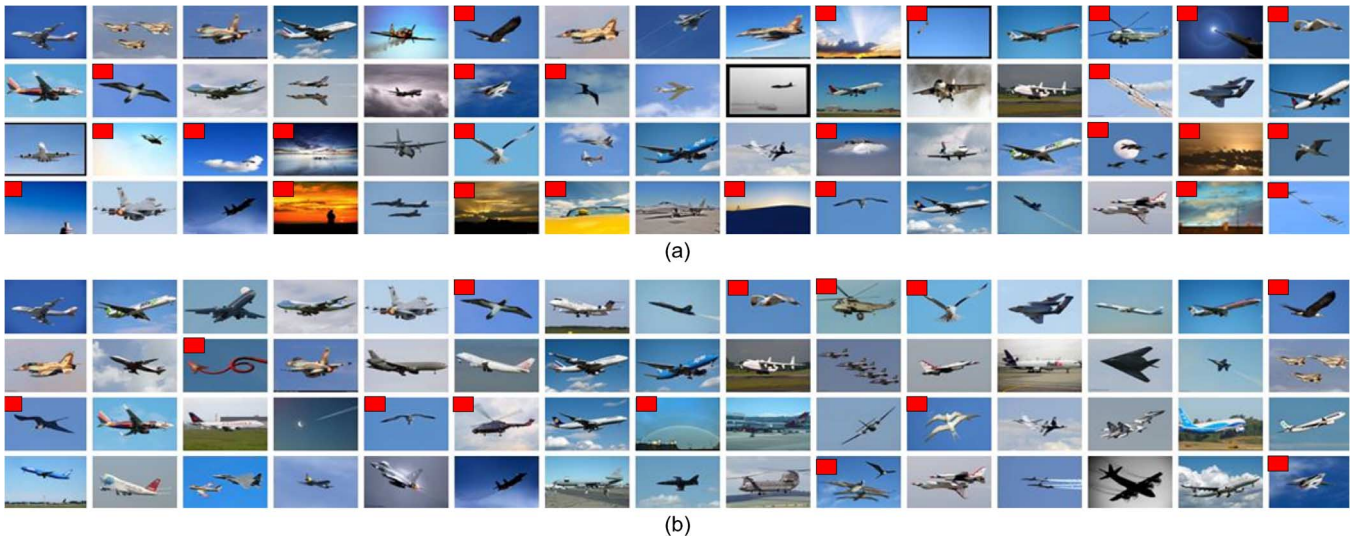


Fig. 6. Top-60 plane classification results ranked by SVM prediction score using (a) SF and (b) SP labels, respectively. Images are ordered from left to right and top to bottom. False-alarms are marked with red rectangular (in online version).

mistakes. In this example, by using SP, the AP performance of concept "plane" is significantly improved from 0.172 (SF) to 0.240 (SP).

SP-I also improves both SF and KW significantly, with an MAP of 0.174. Comparing SP-I to SP, there is no clear winner between the two. SP-I is computationally more efficient, but is also more sensitive to parameter setting simply because it involves a few more parameters. In addition, comparing our SP result $(\mathrm{MAP} = 0.176)$ with the classifiers using expert labels $(\mathrm{MAP} = 0.237)$, there is still a performance gap. This is not surprise because the NUS-WIDE dataset was constructed two years ago by crawling images from Flickr. Images downloaded by the same query at different time may be visually quite dissimilar, resulting in a visual gap between our newly crawled images and the official NUS-WIDE testing images, which may affect the concept learning performance.

D. Experiment 3: Cross Dataset Evaluation

In order not to overlook the issue of over-fitting, we now compare the classifiers built based on expert-labeled examples and free samples by our approach *across* different datasets. In addition to NUS-WIDE, in this experiment we also adopt the VOC 2010 and ImageNet30 datasets. We examine the performance of classifiers learned from the training set of one dataset (e.g., NUS-WIDE) on the testing set of another dataset (e.g., VOC and ImageNet). Based on this comparison, we aim to study the generalization capability of classifiers built from automatic sampling of Web images.

We first evaluate the models used in Experiment 2 on the test set of VOC 2010. Results are reported on the eight concepts common to NUS-WIDE and VOC 2010. As shown in Fig. 7, SP/SP-I achieves the best results for most of the concepts. Overall, the learning performance of SP $(\mathrm{MAP} = 0.334)$ is slightly better than that of expert-labeled training set provided by NUS-WIDE $(\mathrm{MAP} = 0.332)$. This verifies our suspicion earlier that expert labels of NUS-WIDE perform good on the dataset itself because of data domain over-fitting. In other words, SP is actually able to produce training sets comparable to

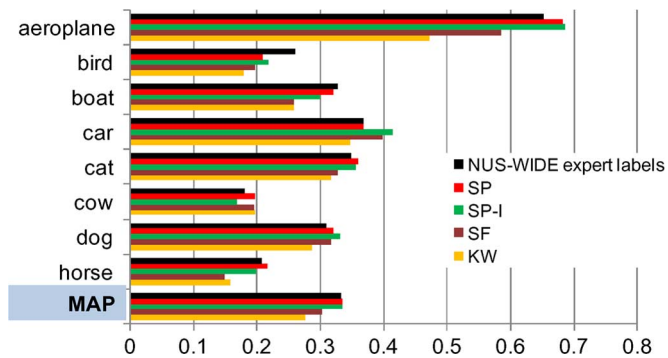


Fig. 7. AP performance of models trained with free Web images on PASCAL VOC 2010 test set.

TABLE IV
NUMBER OF POSITIVE TRAINING SAMPLES IN NUS-WIDE, IMAGENET30, AND VOC 2010

	NUS-WIDE	ImageNet30	VOC 2010
aeroplane	1,584	717	579
bird	2,224	1,063	666
boat	2,477	602	432
car	967	654	1,030
cat	1,425	743	1,005
dog	1,486	802	1,199
horse	1,038	701	425

the expert labeled training sets. Similar observations are noticed from another experiment on the ImageNet30 dataset, which has 30 concepts in common to NUS-WIDE. As can be seen in Fig. 8, SP $(\mathrm{MAP} = 0.288)$ performs better than NUS-WIDE expert labels $(\mathrm{MAP} = 0.258)$. There are 18 out of 30 concepts showing improvement in classification. In addition, similar to the conclusions from Experiment 2, we also observe significant performance improvements from SP/SP-I over SF and KW on both VOC 2010 and ImageNet30 datasets.

We further compare three sets of classifiers (NUS-WIDE model, ImageNet30 model, and VOC model) learned from expert labels of NUS-WIDE, ImageNet30 and VOC 2010 respectively with classifiers based on Web images (SP model).

TABLE V
CROSS DATASET EVALUATION RESULTS. ALL THE MODELS ARE LEARNED WITH A FIXED NUMBER OF TRAINING IMAGES (400 POSITIVE AND 5000 NEGATIVE SAMPLES)

	Test on NUS-WIDE				Test on ImageNet30				Test on VOC 2010			
	NUS-WIDE model	SP	ImageNet30 model	VOC model	NUS-WIDE model	SP	ImageNet30 model	VOC model	NUS-WIDE model	SP	ImageNet30 model	VOC model
aeroplane	**0.222**	0.163	0.179	0.147	0.426	0.371	**0.580**	0.487	0.613	0.597	0.706	**0.732**
bird	**0.150**	0.117	0.059	0.040	0.188	0.179	**0.313**	0.134	0.208	0.180	0.243	**0.388**
boat	**0.157**	0.074	0.079	0.104	0.123	0.103	**0.314**	0.143	0.269	0.255	0.379	**0.534**
car	**0.086**	0.068	0.055	0.034	0.271	0.300	**0.597**	0.333	0.329	0.331	0.450	**0.419**
cat	0.064	**0.098**	0.040	0.025	0.205	0.191	**0.321**	0.225	0.319	0.316	0.353	**0.452**
dog	0.050	**0.067**	0.022	0.024	0.116	0.107	**0.227**	0.091	0.272	0.309	0.276	**0.363**
horse	**0.127**	0.106	0.054	0.041	0.169	0.192	**0.317**	0.222	0.177	0.168	0.251	**0.404**
MAP	**0.122**	0.099	0.070	0.059	0.214	0.206	**0.381**	0.234	0.312	0.308	0.380	**0.470**

TABLE VI
CROSS DATASET EVALUATION RESULTS. NUS-WIDE, IMAGENET30, AND VOC MODELS ARE LEARNED WITH THE MAXIMUM NUMBER OF TRAINING IMAGES AVAILABLE IN EACH DATASET. SP MODELS ARE TRAINED ON 2000 POSITIVE SAMPLES AND 5000 NEGATIVE SAMPLES

	Test on NUS-WIDE				Test on ImageNet30				Test on VOC 2010			
	NUS-WIDE model	SP	ImageNet30 model	VOC model	NUS-WIDE model	SP	ImageNet30 model	VOC model	NUS-WIDE model	SP	ImageNet30 model	VOC model
aeroplane	**0.276**	0.251	0.210	0.149	0.458	0.447	**0.596**	0.501	0.652	0.681	0.713	**0.741**
bird	**0.196**	0.181	0.074	0.044	0.251	0.211	**0.375**	0.142	0.260	0.209	0.285	**0.400**
boat	**0.193**	0.100	0.082	0.102	0.143	0.135	**0.346**	0.144	0.327	0.320	0.394	**0.531**
car	**0.092**	0.086	0.065	0.058	0.300	0.362	**0.606**	0.399	0.368	0.367	0.462	**0.470**
cat	0.112	**0.138**	0.048	0.040	0.243	0.243	**0.354**	0.265	0.348	0.359	0.378	**0.480**
dog	0.086	**0.096**	0.029	0.036	0.146	0.138	**0.279**	0.107	0.310	0.319	0.315	**0.415**
horse	0.150	**0.187**	0.056	0.052	0.220	0.245	**0.357**	0.232	0.207	0.216	0.300	**0.414**
MAP	**0.158**	0.148	0.081	0.069	0.252	0.254	**0.416**	0.256	0.353	0.353	0.407	**0.493**

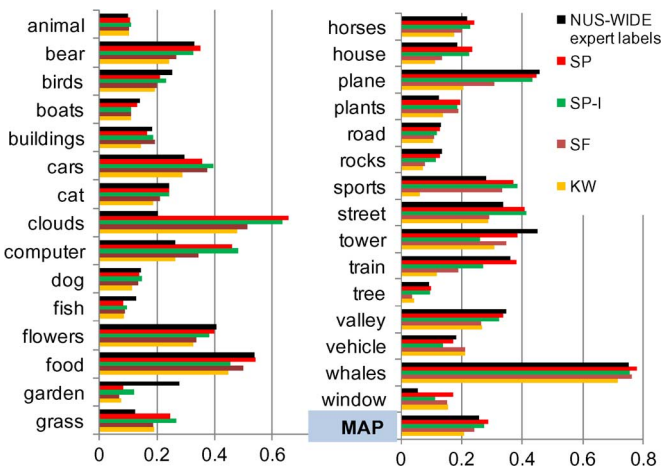


Fig. 8. AP performance of models trained with free Web images on ImageNet30 test set.

Among the three datasets, there are seven common concepts. Since the size of training sets is different across the three datasets (see Table IV), we conduct two sets of cross dataset evaluation experiments. The first one sub-samples the expert-labeled training data and uses a fixed number of examples for all the evaluated models, to make sure results from this experiment are directly comparable. Specifically, all the models are trained using 400 positive samples and 5000 negative samples. For SP, we use 200 images from the target parent node and another 200 pooled from its child nodes.

Results are summarized in Table V, where we can see that within-dataset models always perform the best. This is not surprising at all since images within the same dataset tend to follow similar distribution in feature space. In other words, there is always a domain change across different datasets. This observation is also in line with a recent study in the computer vision community by Torralba and Efros [43]. Although our SP models are not as good as those trained on within-dataset expert labels, they perform consistently well on all the three datasets. Very interestingly, on the NUS-WIDE test set, SP models even outperform both ImageNet30 and VOC models.

The second set of cross dataset evaluation experiments uses the maximum number of examples available in each dataset, aiming to understand the upper limit of the models trained with the expert-labeled images. For SP model, we follow Experiment 2 to use a fixed number of 2000 positive samples and 5000 negative samples. Table VI shows the results. Within-dataset models still exhibit the best overall performance in this experiment, and SP models again performs consistently well on all the three datasets being tested. With more training samples than those used for the results in Table V, this time SP models in Table VI produce better performance which is closer to the best result (from the within-dataset models) over each dataset.

It is worth noting that the VOC/ImageNet models perform better than the NUS-WIDE models on ImageNet (VOC) dataset. The performances of different models on different datasets are also related to the way the datasets were collected. Images in NUS-WIDE were annotated by students from high schools and colleges, while ImageNet and VOC images were annotated by workers on Amazon Mechanical Turk or expert researchers, who have different knowledge background. More importantly, the criteria for characterizing the presence of concepts could result in performance bias. In VOC and ImageNet, each concept is associated with a verbose textual description as guideline for manual annotation. Particularly, ImageNet uses the definitions based on WordNet and Wikipedia. NUS-WIDE adopts a relatively loose instruction. For instance, examples of "car"

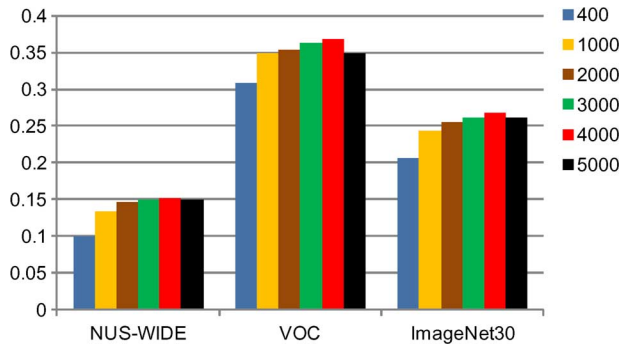


Fig. 9. MAP performance of SP models trained with various numbers (400–5000) of sampled images.

in VOC include cars, vans and large family cars etc., but exclude go-carts, tractors, emergency vehicles and lorries/ trucks etc. In NUS-WIDE, there are diverse examples of "car" such as jeep and fire truck. In addition, images in NUS-WIDE and VOC were downloaded from Flickr only, while ImageNet contains images downloaded from several other search engines like Google. These factors are expected to impact the classification performance, which is however very difficult to be quantitatively measured. Overall there are two major observations from the results shown in Tables V and VI. First, models trained with expert-labeled images tend to over-fit a particular dataset, and often show significant performance degradation when applied to a new dataset. Second, models trained on freely available Web images, with a well-designed sampling method, can perform and generalize fairly well in practice.

Finally, we conduct our last experiment to investigate the classification performance when more training examples are sampled by SP to learn classifiers. Fig. 9 shows the results on three datasets for positive examples ranging from 400 to 5000 per concept. Increasing the number of examples basically leads to gradual improvement in MAP until reaching a level when the sample size is close to 5000. From our observation, the sampled images ranked in the lower part of a list are either irrelevant to the target concept or near-duplicates of other higher ranked examples. Thus, further increasing the sample size may not imply performance improvement, especially for specific concepts. Nevertheless, as social media websites receive a large number of new image upload every day, we expect the general trend of improvement w.r.t the number of samples as long as more relevant and novel images are found in the lower part of the search result list.

V. Conclusion

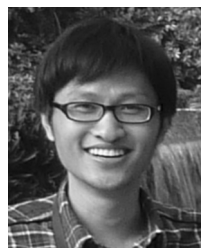
We have presented an efficient approach for sampling and pooling free samples from the Web for training set construction. The modeling of tag lists with SF leads to the effective ranking of sample relevancy, while the hierachical pooling of samples with ontological relatedness enriches the coverage of training sets. Through empirical studies we have verified the merit of seeking samples for free with the goal of maximizing sample relevancy and coverage. More importantly, further cross-dataset evaluation also reveals that by doing so the classifiers learned from free samples can exhibit competitive performance, especially in terms of generalization ability, when comparing with the classifiers learned from expert labeled data. Our work is suitable for sampling Web images that are rich of user tags such as from Flickr. For images that contain limited tags (e.g., from Google and Bing Image Search) or images that tend to be tagged at a higher album-level (e.g., from Picasa), concept learning performance from automatic sampling is expected to be lower.

Currently, our work only considers the collection of positive examples for classifier learning. Future extension includes the revision of current formulation as a unified model for sampling and pooling of both positive and negative examples. The extension shall consider not only the criteria: relevancy, coverage and diversity, but also other factors such as visual, semantic and contextual relatedness. For instance, the examples from sibling concepts are more "negatively" informative as reported in [25] (eg, training "cat" classifier using images of "dog" as negative samples). Nevertheless, adding them in classifier learning may degrade the performance as the "visual context" of sibling concepts could be similar. Therefore, how to find a trade-off among different factors for maximizing the performance of classifiers learnt with free positive and negative samples is an interesting problem to be investigated.

References

[1] G. Quénot, B. S. F. Thollard, and S. Ayache, "TRECVID 2011 collaborative annotation," 2011. [Online]. Available: http://mrim.imag.fr/tvca/
[2] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge 2010 (VOC2010) results," 2010. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html
[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 20–25, 2009, pp. 248–255.
[4] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, "To search or to label," in *Proc. ACM MIR*, 2006, pp. 249–258.
[5] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
[6] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *IJCV*, vol. 73, no. 2, pp. 213–238, 2007.
[7] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 42–53, Jan. 2010.
[8] S.-F. Chang *et al.*, "Columbia University/VIREO-CityU/IRIT TRECVID 2008 high-level feature extraction and interactive video search," in *NIST TRECVID Workshop*, 2008.
[9] C. G. M. Snoek *et al.*, "The MediaMill TRECVID 2010 semantic video search engine," in *Proc. NIST TRECVID Workshop*, 2010, pp. IV-1213–IV-1216.
[10] R. Yan and M. R. Naphade, "Semi-supervised cross feature learning for semantic concept detection in video," in *Proc. CVPR*, 2005, pp. 657–663.
[11] Q. Tian, J. Yu, Q. Xue, and N. Sebe, "A new analysis of the value of unlabeled data in semi-supervised learning for image retrieval," in *Proc. ICME*, Jun. 30, 2004, no. 1019, p. 1022.
[12] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proc. ACM MM*, 2007, pp. 188–197.
[13] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, Tech. Rep. 7694, 2007. [Online]. Available: http://authors.library.caltech.edu/7694
[14] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. ICCV*, 2005, pp. 1800–1807.
[15] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *IJCV*, vol. 77, no. 1–3, pp. 157–173, 2008.
[16] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. ACM MIR*, 2008, pp. 39–43.
[17] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," presented at the CIVR, New York, NY, 2009.

[18] L. Kennedy and A. Hauptmann, "Lscom lexicon definitions and annotations," Columbia Univ., New York, NY, ADVENT Tech. Rep. 217-2006-3, 2006.
[19] A. T. Setz and C. G. M. Snoek, "Can social tagged images aid concept-based video search," in *Proc. ICME*, 2009, pp. 1460–1463.
[20] A. Ulges, C. Schulze, M. Koch, and T. M. Breuel, "Learning automatic concept detectors from online video," *CVIU*, vol. 114, no. 4, pp. 429–438, 2010.
[21] J. Wang, Y.-G. Jiang, and S.-F. Chang, "Label diagnosis through self tuning for web image search," in *Proc. CVPR*, 2009, pp. 1390–1397.
[22] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," in *Proc. ICCV*, 2007, pp. 754–766.
[23] L.-J. Li and L. Fei-Fei, "OPTIMOL: Automatic object picture collection via incremental model learning," *IJCV*, vol. 88, no. 2, pp. 147–168, 2009.
[24] X.-R. Li and C. G. M. Snoek, "Visual categorization with negative examples for free," in *Proc. ACM MM*, 2009, pp. 661–664.
[25] X.-R. Li, C. G. M. Snoek, M. Worring, and A. W. Smeulders, "Social negative bootstrapping for visual categorization," in *Proc. ICMR*, 2011, p. 12.
[26] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.
[27] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proc. ICCV*, 2005, pp. 1816–1823.
[28] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 754–766, Apr. 2011.
[29] J. Fan, Y. Shen, N. Zhou, and Y. Gao, "Harvesting large-scale weakly-tagged image databases from the web," in *Proc. CVPR*, 2010, pp. 802–809.
[30] A. Ulges, C. Schulze, D. Keysers, and T. Breuel, "Identifying relevant frames in weakly labeled videos for training concept detectors," in *Proc. CIVR*, 2008, pp. 9–16.
[31] A. Ulges, D. Borth, and T. Breuel, "Visual concept learning from weakly labeled web videos," in *Video Search and Mining*. New York: Springer-Verlag, 2010.
[32] D. Borth, A. Ulges, and T. Breuel, "Relevance filtering meets active learning: Improving web-based concept detectors," in *Proc. MIR*, 2010, pp. 25–34.
[33] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain, "Image annotation by nn-sparse graph-based label propagation over noisily tagged web images," *Proc. ACM TIST*, vol. 2, no. 2, p. 14, 2011.
[34] D. Liu *et al.*, "Tag ranking," in *Proc. ACM WWW*, 2009, pp. 351–360.
[35] X.-R. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.
[36] L. Kennedy, M. Slaney, and K. Weinberger, "Reliable tags using image similarity: Mining specificity and expertise from large-scale multimedia databases," in *Proc. 1st Workshop Web-Scale Multimedia Corpus*, 2009, pp. 17–24.
[37] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba, "Semantic label sharing for learning with many categories," in *Proc. ECCV*, 2010, pp. 762–775.
[38] S. Zhu, G. Wang, C.-W. Ngo, and Y.-G. Jiang, "On the sampling of web images for learning visual concept classifiers," in *Proc. CIVR*, 2010, pp. 50–57.
[39] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2000.
[40] W. Zhibiao and M. Palmer, "Verb semantic and lexical selection," in *Proc. ACL*, 1994, pp. 133–138.
[41] Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang, "Semantic context transfer across heterogeneous sources for domain adaptive video search," in *Proc. ACM MM*, 2009, pp. 155–164.
[42] R. L. Cilibrasi and P. M. B. Vitányi, "The google similarity distance," *IEEE Trans. Knowledge Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.
[43] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, Jun. 20–25, 2011, pp. 1521–1528.



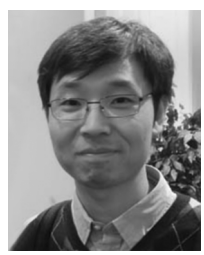
Shiai Zhu received the B.Eng. degree in electronic information engineering from the Civil Aviation University of China, Tianjin, China, in 2005, and the M.Eng. degree in signal and information processing from the University of Science and Technology of China, Hefei, China, in 2008, and is currently pursuing the Ph.D. in computer Science at the City University of Hong Kong, Kowloon, Hong Kong.

He was a Software Engineer at the Huawei Technologies Company, Ltd., Shanghai, China, from 2008 to 2009. His research interests include multimedia analysis and multimedia information retrieval.

Chong-Wah Ngo received the and MSc and B.Sc. degrees in computer engineering from the Nanyang Technological University, Singapore, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2000.

He is an Associate Professor in the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. Before joining City University, he was a Postdoctoral Scholar in Beckman Institute, University of Illinois at Urbana–Champaign (UIUC). He was also a Visiting Researcher of Microsoft Research Asia. His recent research interests include large-scale multimedia information retrieval, video computing, and multimedia mining.

Dr. Ngo is currently an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA. He is also the program co-Chair of ACM Multimedia Modeling (MMM) 2012 and the International Conference on Multimedia Retrieval (ICMR) *2012*. He served as the Chairman of ACM (Hong Kong Chapter) during 2008–2009.



Yu-Gang Jiang received the Ph.D. degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in 2009.

During 2008–2011, he was with the Department of Electrical Engineering, Columbia University, New York, NY, first as a Visiting Scholar and later as a Postdoctoral Research Scientist. He is currently an Associate Professor of computer science at Fudan University, Shanghai, China. He has authored more than 30 papers in these fields. He is an active participant of the Annual NIST TRECVID Evaluation and has designed a few top-performing video retrieval systems over the years. His research interests include multimedia retrieval and computer vision.

Dr. Jiang has served on the technical program committees of many international conferences, and is a Guest Editor of a forthcoming special issue of the IEEE TRANSACTIONS ON MULTIMEDIA on socio-video semantics.