On the Sampling of Web Images for Learning Visual Concept Classifiers

Shiai Zhu[†], Gang Wang[†‡], Chong-Wah Ngo[†], Yu-Gang Jiang[§]
[†]Dept of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[‡]School of Computer Science, Fudan University, Shanghai, China
[§]Dept of Electrical Engineering, Columbia University, New York, NY, USA
shiaizhu2@student.cityu.edu.hk, wanggang_sh@hotmail.com
cwngo@cs.cityu.edu.hk, yjiang@ee.columbia.edu

ABSTRACT

Visual concept learning often requires a large set of training images. In practice, nevertheless, acquiring noise-free training labels with sufficient positive examples is always expensive. A plausible solution for training data collection is by sampling the largely available user-tagged images from social media websites. With the general belief that the probability of correct tagging is higher than that of incorrect tagging, such a solution often sounds feasible, though is not without challenges. First, user-tags can be subjective and, to certain extent, are ambiguous. For instance, an image tagged with "whales" may be simply a picture about ocean museum. Learning concept "whales" with such training samples will not be effective. Second, user-tags can be overly abbreviated. For instance, an image about concept "wedding" may be tagged with "love" or simply the couple's names. As a result, crawling sufficient positive training examples is difficult. This paper empirically studies the impact of exploiting the tagged images towards concept learning, investigating the issue of how the quality of pseudo training images affects concept detection performance. In addition, we propose a simple approach, named semantic field, for predicting the relevance between a target concept and the tag list associated with the images. Specifically, the relevance is determined through concept-tag co-occurrence by exploring external sources such as WordNet and Wikipedia. The proposed approach is shown to be effective in selecting pseudo training examples, exhibiting better performance in concept learning than other approaches such as those based on keyword sampling and tag voting.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Measurement, Experimentation

Keywords

Concept Detection, Web Images, Sampling

1. INTRODUCTION

Visual concept detection is fundamentally a classification task that determines whether a multimedia unit (e.g., image) is relevant to a given target concept. Specifically, classifiers (e.g., SVM) are trained with training examples, and the learnt classifiers are employed for concept annotation. A critical step along this process is the acquisition of sufficiently large amount of quality training data for concept learning. The acquisition, nevertheless, is not a trivial process. Labeling TRECVID 2009 dataset, for instance, requires collaborative efforts from about 40 research teams to manually annotate 43,616 shots for 12 concepts [18]. Such a labor-intensive process will become extremely difficult for the ultimate aim of labeling thousands of visual concepts.

On the other hand, with the popularity of social media, there are more and more digital images annotated with user tags and comments on the web. For example, it is reported that there are four billions of images on the Flickr web site. Automatic sampling of these weakly labeled web images for concept learning thus appears as a natural way of replacing expensive manual labeling. Such efforts include the recent works in [8, 15]. In [8], for the purpose of filtering noisy tagged images, semi-supervised learning is adopted to collect web images similar to expert-labeled images for cross-domain concept learning. In [15], in view that as high as 90% of manual labeling efforts are spent on identifying negative samples, concept learning is conducted by direct collection of negative samples from user-tagged images, together with expert-labeled positive examples. Despite of these efforts, in general, sampling of noise-free training samples, especially the positive samples, from the web for effective classifier learning remains an issue not fully understood. How such weakly labeled examples affect learning, and the proper way of acquiring training examples for free, are yet to be addressed. The study in [13] reveals that the user-supplied tags are imprecise and only 50% of tags are related to images.

Gang Wang is the corresponding author.

CIVR '10, July 5-7, Xi'an, China

Making photos accessible to public is only one of the tagging motivation from users. As indicated by [6], there are roughly eight categories of tagging motivation including opinion expression, attraction of attention, and self-presentation. Under such scenario, user tags can be personalized, and more importantly, may not be helpful for visual search.

The challenges of sampling web images for concept learning can be briefly summarized as follows. First, user-tags tend to be subjective and personalized. Different users may focus on different aspects of an image and thus provide a variety of tags which are not necessarily content-related. For example, Figure 1(a) is an image about child. However, the concept "child" is not tagged. Instead, context-related tags such as "barossa" and "valley", indicating the place and location where the photo was taken, are tagged. Collecting images as 1(a) for training concepts such as "valley" will lead to ineffective classifier learning. Second, user tags can be ambiguous. The photo in 1(b) contains the tag "bear". While "bear" is commonly referred to as an animal, it also has another word sense: "a surly, uncouth, burly, or shambling person". In other words, while the tag "bear" for 1(b) is content-related, certain degree of word sense disambiguation is required to differentiate "bear" in this photo as a person from an animal. Third, tag list is often incomplete. For instance, in 1(c), while the concepts such as "sea", "sky" and "water" are tagged, the concept "beach" is missing. In this case, sampling training images merely based on keyword can result in low recall of positive samples.

Figure 1: Examples of web images with tags.

This paper studies two issues: 1) how to sample pseudo positive/negative training data from web images, and 2) the quality of training data towards concept classifier learning. Given a web image with a tag list, we propose an approach to predict the "Semantic Field" of the image. Semantic Field [12] was originally proposed to capture a more integrated relationship among the entire set of words. In our work, we consider four different cases of examples, as shown in Figure 1. In 1(a), the image will not be sampled for training concept "valley" since the surrounding tags such as "boy" and "laugh" do not support the existence of "valley" in the image. In 1(b), the concept "bear" could possibly be disambiguated as not related to "animal" – the most common sense of "bear", by investigating other tags such as "beach" and "celebrity". In 1(c), though concept "beach" is missing, the image will still be sampled for learning "beach" since surrounding tags such as "sea", "sky" and "sand" jointly suggest the concept of "beach" in the image. In 1(d), the image could possibly be ranked higher for learning concept "dog" since the tags "dog", "animal", "pet" and "puppy" give clue that "dog" is the major highlight of the image. The significance of user tags towards a target concept can be modeled from three different sources: web image copus, Wordnet and Wikipedia. In brief, different from the direct matching of keywords and tags, we consider tags of an image collectively to predicting underlying semantic field. Ideally, the semantic field can highlight the major visual concepts in images while providing an effective means for ranking and sampling pseudo-training images for a given concept.

The remaining sections are organized as follows. Section 2 reviews the related work about concept learning, tag quality refinement, and sampling of pseudo training examples. Section 3 describes the proposed work on modeling the semantic field of concepts for image ranking and sampling. Section 4 presents our empirical study and simulation on the effect of training data quality towards classifier learning. Comparison among different approaches is also given to investigate the performance of the proposed approach based on semantic field. Finally Section 4 concludes the findings in this paper.

2. RELATED WORKS

2.1 Concept Learning

In multimedia community, numerous efforts have been devoted to concept classifier learning. Naphade and Smith [17] surveyed the state-of-the-art systems and pointed that most of them adopted supervised learning approaches for semantic concept detection. Such learning approaches estimate a function for all possible input values. This implies the availability of good quality training data, which ideally includes most typical types of the data in the test set. In other words, supervised learning expects that the data distribution of training examples is close enough to that of testing data such that learning can be effective. The detection performance may degrade significantly if this condition could not be satisfied.

To address this problem, Yan and Naphade [24] proposed a multi-view semi-supervised cross-feature learning approach. Initially one classifier from each view is learnt 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. [7] pinpointed that unlabeled data helps only if labeled and unlabeled data are from the same distribution in a semi-supervised learning framework. Otherwise, detection performance may degrade. Qi et al. [4] and Wang et al. [22] proposed transductive learning methods to infer unlabeled test data by finding related labeled training data via a clustering method. Tong et al. [21] conducted active learning to capture more related training examples through relevance feedback. Despite these efforts, the gap between training and testing data commonly exists and adversely impacts the effectiveness of classifier. Intuitively, the gap can be bridged by acquiring a sufficiently number of training examples. In view that manual labeling is expensive, automatic sampling of weakly tagged but heterogeneous training data from the web becomes a timely issue to study.

2.2 Quality of User Tagging

Ames and Naaman [1] analyzed the motivations of annotation in mobile and online media. They found that the

these include both personal and social purposes. Bischoff et al. [6] provided the tag distributions in three tagging environments. The study indicated that only 45%-60% tags can be used to enhance search experience. In the view of aiding in searching, other tags such as tagging for owners or self references belong to the noise.

Kennedy et al. explored the trade-off in acquiring training data by automated web image search as opposed to human annotation [13]. The study, based on concepts in consumer photos, indicated that concept classifiers learnt from manual annotation generally outperform those learnt from training examples acquired from image search. However, for concepts which are visually consistent across domains, there is no apparent performance difference between using human annotation and noisy web images. This suggests that manual annotation is not necessary for this group of concepts. On the other hand, for concepts which are visually diverse across different domains, the performance is equally worse regardless of using human annotation or search result. This also gives clue that human annotation may not be necessary for this category of images. The study showed that there are mainly two cases where human annotation will be helpful. These include concepts which have many view angles but are visually consistent across domains, and concepts which lack of coherence across domains but share visual consistency between training and testing data.

Tag refinement is one effort aiming for improving tagging quality. Liu et al. [3] adopted random walk over a tag similarity graph to refine the relevance scores. Li et al. [16] proposed a neighbor voting algorithm to learn tag relevance by accumulating votes from visual neighbors. While these works indicated that content relevant tags can be somewhat predicted from noisy user tags and visual similarity, these tags were also considered seperately and there is no effort yet showing the effect of using the images under tag refinement for concept learning. One simulation study was conducted by [20] using images with *manually* disambiguated tags. It was report that training images with higher quality will improve the performance of most concept classifiers either in within-domain or cross-domain scenarios.

2.3 Collecting Pseudo Training Examples

Acquiring training images from the web for various purposes including concept classifier learning has recently captured numerous research attentions [20, 5, 14, 2]. One common technique is to start with tag-based visual search such as keyword matching or query expansion for collecting training samples. The initial search list is then utilized directly for classifier learning [20, 2]. More advanced techniques include the refinement of search list with machine learning techniques. In [19], a two-step approach was proposed. Firstly a Bayes posterior estimator is trained on the surrounding metadata of images to rerank the initial list. Then the top ranked images are used to learn a SVM classifier to further refine the ranking. In [5], semi-supervised learning is adopted to harness tagged and untagged images simultaneously to alleviate the effect of noisy tagging. In [14], an iterative concept learning and image collecting framework was provided. Starting from a small number of training images, a model is trained for each class to sample the text search results. The newly collected images serve as additional training data for refining the original learnt classifier. By iterative learning, it is expected that the robustness of the classifiers will be improved, while more quality training images can be acquired. These refinement techniques, however, are computationally very expensive [5, 14] and therefore unscalable for large scale data sets [5].

While acquiring noise-free positive examples is difficult, obtaining negative samples appears feasible even with simple heuristics. Yan et al. [23] proposed an approach to collect negative samples by exploiting the most irrelevant images via content-based retrieval. On the other hand, Liu et al. [3] adopted text-based analysis. Given a concept, synonyms and descendant related words are expanded to the concept. Negative training samples are collected by eliminating images tagged with the related words. In addition, Li et al. [15] empirically shown that replacing expert-labeled negative examples with social tagged images for concept learning only results in slight loss of detection performance.

3. MODELING SEMANTIC FIELD

In this section, we introduce a method called Semantic Field (SF) to measure the relevance of tags to concepts. Semantic Field is to capture the semantics of a set of words [12]. The basic idea is that the meaning of a word is dependent partly on its relation to other words. This implies that mining the semantics of a word list needs collective analysis, rather than interpreting each word in the list individually. Translating this to our application, each tag list of an image basically carries one semantic field. In this section, we propose a probabilistic model to describe the association between a target concept and a tag list. For efficiency, we also build a dictionary for each target concept. The dictionary is composed of a list of reference words that depicts the semantic field of a target concept. We explore this dictionary to evaluate the degree of association between tag lists and the target concept. Different from tag refinement [3, 16] which ranks tags of an image, we employ SF to rank the pseudo training images of a concept.

3.1 A Probabilistic Model

Intuitively, some tags in a list are correlated to each other. A tag list can be integrally considered as carrying a Semantic Field. Denote C_x as a target concept, and $SF =< T_1, T_2, \ldots, T_n >$ as the tag list of an image I containing n tags. The probability of C_x in I is defined as:

$$P(C_x|SF) = \frac{P(SF|C_x) \times P(C_x)}{P(SF)} \quad (1)$$

where the prior probability $P(C_x)$ can be viewed as a constant and therefore can be ignored for the purpose of ranking based on $P(C_x|SF)$. We approximate $P(SF|C_x)$ using $P(SF) * (\sum P(T_i|C_x)/n)$, and Equation (1) can then be rewritten as:

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

where $P(T_i|C_x)$ donates the likelihood of observing a tag T_i given a concept C_x.

3.2 Learning SF from Multiple Sources

To estimate $P(T_i|C_x)$ in Equation (2), we consider both domain and general knowledge. For domain knowledge, we utilize the co-occurrence statistics of tag lists which can be computed from any web image corpus. For general knowledge, we exploit Wordnet and Wikipedia for information

Table 1: Statistics of 81 labeled concepts in NUS-WIDE corpus.

Categories	Number of concepts	Average number of training samples	Average tag frequency
people	4	8494	916
objects	33	1956	1405
scene/location	33	6015	2436
event/activities	9	366	645
program	1	1104	758
graphics	1	40	212

inference. Combining different knowledge sources, we have

$$P(T_i|C_x) = P_{wd}(T_i|C_x) \times P_{wiki}(T_i|C_x) \times P_{co}(T_i|C_x) \quad (3)$$

where $P_{wd}(T_i|C)$ is the probability of observing tag T_i after querying Wordnet with concept C_x. Similarly $P_{wiki}(T_i)$ and $P_{co}(T_i)$ are the probabilities inferred from Wikipedia and co-occurrence statistics of tags in our Flickr corpus respectively.

The three knowledge sources provide different aspects of information. WordNet models the relatedness of words as a graph and lists the multiple senses of a word. Wikipedia provides less structural information but gives comprehensive description of concepts. While both WordNet and Wikipeda are text based, tag co-occurrence computed from tag lists of web images provides an objective view of tag statistics in the domain of visual data. Based on different natures of information sources, we compute Equation (3) as following:

$$P_{wd}(T_i|C_x) = \frac{\#(T_i, C_x)}{\#(C_x)} \approx \frac{\#(T_i)}{\#(words)_{wd}} \quad (4)$$

$$P_{wiki}(T_i|C_x) = \frac{\#(T_i, C_x)}{\#(C_x)} \approx \frac{\#(T_i)}{\#(words)_{wiki}} \quad (5)$$

$$P_{co}(T_i|C_x) = \frac{\#(T_i, C_x)}{\#(C_x)} \quad (6)$$

where $\#(T_i, C_x)$ is the number of co-occurrences between tag T_i and concept C_x, and $\#(C_x)$ is the term frequency of concept C_x. We approximate Equation (4)-(5) by counting the term frequency of T_i, $\#(T_i)$, against the number of non-stop words, $\#(words)_{wd}$ and $\#(words)_{wiki}$, respectively from the WordNet and Wikipedia pages which describe concept C_x. For WordNet, we choose the most common sense to describe target concept, while for Wikipedia, we download the related page and compute $P_{wiki}(T_i|C_x)$. In general, $\#(words)_{wiki} > \#(words)_{wd}$. For tag co-occurrence, the denominator $\#(C_x)$ is simply the number of images tagged with concept C_x.

In practice, the number of co-occurrence $\#(T_i, C_x)$ can be equal to zero. We employ add-one smoothing technique [12] as following to deal with the problem:

$$P = \frac{\#(T_i, C_x) + 1}{\#(C_x) + 1} \quad (7)$$

3.3 Dictionary Construction for Image Ranking

Online querying different information sources for computing Equation (2) can be time consuming. For efficiency consideration, we offline build a dictionary for each target concept. A large pool of image tags is first crawled from image search engines. Given a target concept C_x, Equation (3) is evaluated to rank the tags in the pool according to their probability scores. The dictionary of C_x is then formed by including the top-k ranked tags. In our current implementation, k is set to 200 in order to reduce computational cost. From our observation, the scores of the tags ranked after 200 are mostly very small.

The dictionary basically captures the set of words related to the target concept. With the dictionary, given an image and its tag list, Equation (2) can be efficiently computed by dictionary look-up and score averaging. After that, given a target concept, the set of candidate images can be easily ranked, and pseudo positive training samples are then selected according to their scores based on Equation (2).

4. EXPERIMENTS

We split the experiments into two major parts. The first part examines the quality of training samples in affecting the concept detection performance, in which we assume the number of positive samples to be known and randomly choose a set of negative samples. The second part of experiments considers a more realistic scenario in which the number of positive samples is unknown and we therefore select a fixed number of top ranked images as positive set. We compare our approach to several existing techniques including neighbor voting [16] and keyword-based image sampling [20, 2].

4.1 Dataset and Performance Evaluation

We use the recently released web image dataset, NUS-WIDE [9], for performance evaluation. The dataset includes 269,648 images crawled from Flickr, with a total of 5,018 unique tags. Images in NUS-WIDE corpus are manually labeled to provide ground-truth for 81 concepts. This forms 161,789 images for training, and 107,859 images for testing. The 81 concepts are divided into six categories: people, objects, scene or location, event or activities, program and graphics. Table 1 lists the number of concepts, the average number of training samples under each category and the average tag frequency. Tag frequency is the number of training images tagged by users with the target concept. As indicated in Table 1, there is a large difference between the number training images (ground-truth) labeled by human expert and the number images tagged with target concepts by the web users. In the experiment, we conduct testing for 81 concepts based on this subset of NUS-WIDE corpus.

For concept classifier learning, we adopt similar setting as VIREO-374 [10]. For each concept, three SVM classifiers are trained separately based on bag-of-visual-words (BoW), grid-based color moment and wavelet texture respectively. In BoW, local keypoints are randomly sampled from training examples and a visual dictionary of size 500 is constructed. Soft weighting [11] is employed to map multiple keywords to a keypoint, and this forms a BoW of 500 dimensions for each image. For color moment, each image is partitioned into 5×5 grids, and the first three moments are computed on Lab color space over each grid. Concatenating the features from all grids forms a vector of 255 dimensions for each image. Similarly for wavelet texture, each image is divided into 3×3 grids, and each grid is represented by the variances in 9 Haar wavelet sub-bands. This forms a feature vector of 81 dimensions. The raw outputs from the three SVM classifiers are then converted to posterior probabilities using Platt's method. The probabilities are combined as a score,

which indicates the confidence of detecting a concept in an image, by average fusion.

We employ average precision (AP), over the rank list of $N = 107,859$ testing images, to assess the performance of concept detection. Denote R as the total number of true positives (relevant images) in the testing dataset, and R_j as the number of true positives for top-j retrieved images, AP is defined as $AP = \frac{1}{R}\sum_{j=1}^{N} I_j \times \frac{R_j}{j}$, where $I_j = 1$ if the image ranked at j^{th} position is relevant, and $I_j = 0$ otherwise. By averaging the APs of all concepts being tested, mean AP (or MAP) is obtained and used to assess the overall performance.

4.2 Quality of Training Samples

In this section, we examine the quality of positive training samples in affecting concept learning, with an assumption that the number of positive samples for each concept is known, according to the ground-truth annotations. The aim is to provide a fair evaluation when comparing concept detection performance using sample selection approaches to that directly using the ground-truth. In NUS-WIDE, the number of positive training samples varies a lot from concept to concept, depending on their popularities. For instance, popular concept such as "sky" has as many as 44,255 positive samples, while rare concepts such as "map" only has 40 positive samples.

We first compare the performances of the proposed approach (named as Semantic Field) to an oracle setting (Oracle) and a keyword-based sampling method (Keyword). In oracle setting, we use the original training samples provided by NUS-WIDE corpus for learning the 81 concepts. In other words, Oracle should provide the best possible detection performance since all training samples are manually labeled. For Semantic Field and Keyword, no ground-truth labels are given. Instead, the positive training samples are generated by ranking the 161,789 training images in NUS-WIDE by one of both methods. For Semantic Field, a dictionary, as presented in Section 3.3, is constructed for each concept from the unlabeled training images. Then, given a training image, Equation (2) is applied to assign a concept relevancy score for the image. The training images are then ranked according to the scores, and the top-k ranked images are finally used for training concept classifiers. Note that in this experiment the value of k is assumed known for each classifier. For Keyword, a similar setting is used. The training set is prepared by collecting images which are tagged with the target concept. The images are ranked according to the output list sorted by Flickr search engine, and the top-k list are used for classifier learning.

We use the same set of negative samples for all the tested methods in this section. For each concept, we randomly select 5,000 negative samples from the training set after excluding the images which are considered as positive samples by any of the three methods.

4.2.1 *Performance of Concept Classifiers*

Table 2 lists the performances of the three tested approaches. Oracle achieves a MAP of 0.2223 under ideal setting, followed by Semantic Field with $MAP = 0.1660$, and Keyword with $MAP = 0.1242$. Compared with Keyword, Semantic Field can achieve an overall improvement of 33.8%. The improvement is consistently observed in all the six categories of concepts. From our analysis, the performance of Keyword is highly dependent on the correctness of tags provided by users. Referring to Table 1, there is an average of 916 images being tagged with concepts related to the category "people". However, considering the average of 8,494 true positive images being manually labeled, the recall rate is only 10.7%. In other words, by simply matching tags to target concepts, Keyword pays the risk that a large number of positive samples will not be recalled for classifier training. Noisy tags are frequently observed in the categories "events/activities" and "graphics". For instance, among the 645 images tagged with concepts related to "events/activities", 50% of them is regarded as "wrongly" tagged by manual labeling. As a consequence, the training samples collected by Keyword are contaminated by excessive number of false positives for certain concepts such as "soccer", with only 93 true exemplars among the 456 tagged images.

Table 2: Performance of concept detection using different sets of training examples.

Categories	MAP		
	Oracle	Semantic Field	Keyword
people	0.2096	**0.1218**	0.0967
objects	0.1960	**0.1625**	0.1299
scene/location	0.2845	**0.2136**	0.1568
event/activities	0.1101	**0.0522**	0.0211
program	0.1353	**0.0661**	0.0241
graphics	0.1800	**0.0209**	0.0001
all concepts	0.2223	**0.1660**	0.1240

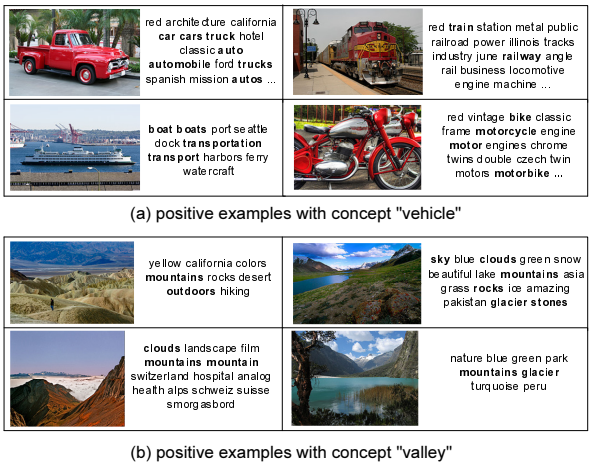

Figure 3: Examples of images which are not tagged with target concepts but correctly labeled by Semantic Field as positive samples.

Semantic Field, in contrast, can deal with both problems quite effectively to certain extent. Figure 3 shows few examples of images which are correctly labeled as positive samples even though the target concept is not tagged by users. In 3(a), the images are tagged with "car", "train" and "boat", but not "vehicle". Based on inference from three information sources, especially the WordNet and Wikipedia, Semantic Field correctly predicts the existence of concept "vehicle" for these images. As a result, the AP of "vehicle" detector is significantly boosted to 0.4300 compared to Keyword whose AP is 0.2480. Similarly in 3(b), the images have several attracting regions and different users provided tags to high-

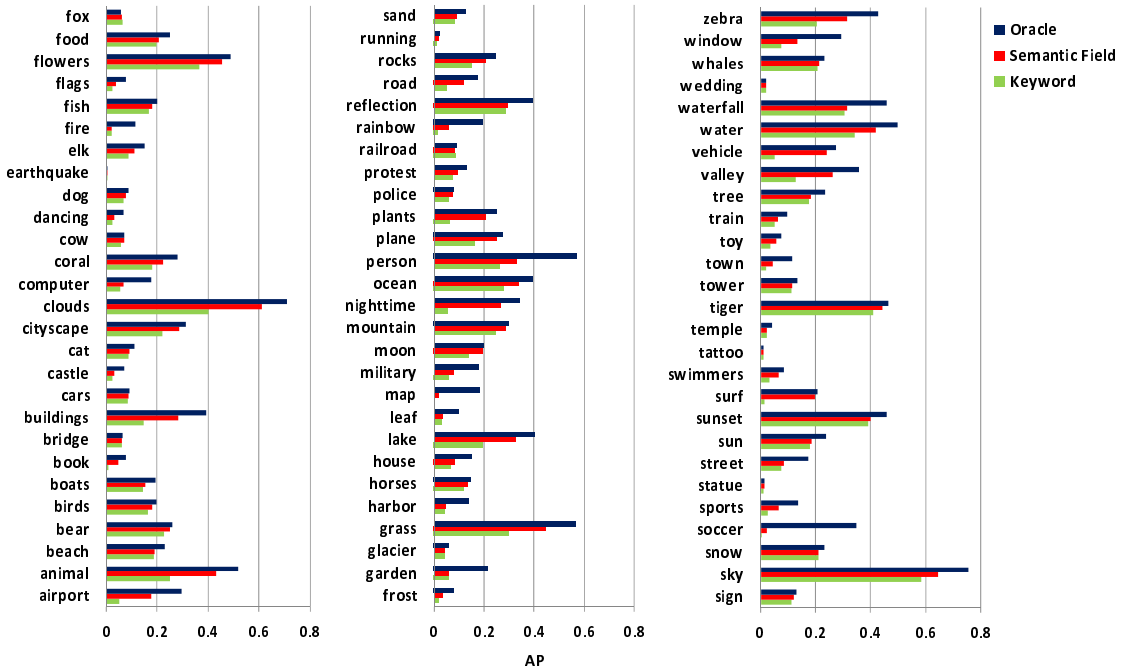

Figure 2: AP comparison of the 81 concepts using the three methods

light different parts of attentions. The tag list of an image, in general, is incomplete. By Semantic Field, nevertheless, the target concept of "valley" can be correctly predicted by tags such as "mountain", "sky" and "clouds". Figure 4 shows a few more examples of images which are misleadingly tagged with "bird". Semantic Field successfully eliminates these samples as positive examples, as the surrounding tags such as "art", "drawing" and "cars" do not show strong enough consistency with the semantic field of concept "bird".

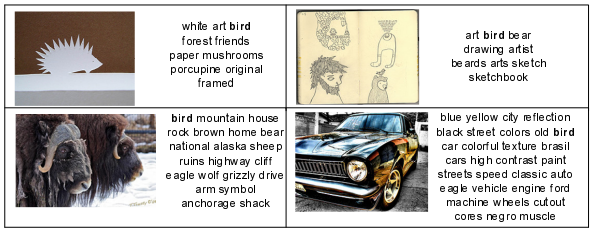

Figure 4: Examples of images which are misleadingly tagged with "bird", but successfully marked as negative samples by Semantic Field.

Figure 2 further details the AP of 81 concepts by the three tested approaches. Almost all the 81 concept classifiers using Semantic Field can achieve better performance than keyword based approach, especially for generic concepts such as "clouds", "animal" and "nighttime". For some of the rare concepts such as "tatoo" and "earthquake", Semantic Field is less effective, since in these cases, robust classifiers are difficult to train even using manually labeled samples. Overall, there is still a performance gap between Semantic Field and Oracle. For concepts such as "leaf" with large intra-class variation, there are insufficient and erroneous training samples. When the amount of noisy images surpasses the correct samples, effective learning of classifiers becomes difficult. On the other hand, for concepts such as "moon" and "snow", we achieve near-optimal. This is mainly due to the fact that the training images of these concepts are visually consistent. Even if the pseudo training samples are contaminated with noises, the classifiers can still be quite effectively learnt.

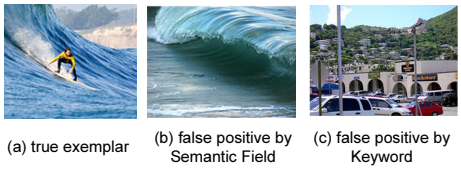

Figure 5: True and false positive examples for concept "surf".

It is interesting to note that some noisy samples are indeed helpful. Figure 5 shows some pseudo training samples for the concept "surf". While Semantic Field samples some images tagged with "ocean" such as 5(b) as training examples of "surf", these images provide contextual cues and are useful for learning concept "surf". Compared to the false positives such as 5(c) sampled by Keyword, Semantic Field has better capability in selecting relevance of samples to a concept.

4.2.2 *Effect of Noise Level*

Table 3 details the quality of pseudo samples generated by Semantic Field and Keyword. We use two measures, noise level (NL) defined in [9] and MAP, to assess the quality of relevance ranking by both methods. NL is defined as $NL = 1 - F1$, where $F1$ is defined as

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

which takes into account the precision and recall of training images. Precision measures the proportion of pseudo

training examples which are correctly labeled, while recall measures the fraction of true positives which are included as training examples. The mean NL (or MNL) is further computed by averaging NL of different concepts. Note that the list of pseudo training images is ranked according to the probability of a concept in the images. Thus, we also use MAP to assess ranking of Semantic Field and Keyword. Basically, AP considers the ranking of training examples, while NL simply assesses the percentage of true samples. Training set with lower MNL value contains more positive samples and less noise, and the set with higher value of MAP ranks more true positives than noises at the top of list.

Table 3: Quality of pseudo training samples measured via MAP (the higher, the better) and MNL (the lower, the better).

Categories	MAP		MNL	
	Semantic Field	Keyword	Semantic Field	Keyword
people	**0.6610**	0.5360	**0.4940**	0.5530
objects	**0.6440**	0.5740	**0.4700**	0.5200
scene/location	**0.5700**	0.5180	**0.5700**	0.6400
event/activities	**0.4380**	0.3650	**0.6530**	0.7000
program	**0.7650**	0.2790	**0.3620**	0.8220
graphics	**0.1510**	0.0000	**0.8750**	1.0000
all concepts	**0.5870**	0.5150	**0.5360**	0.6030

As indicated in Table 3, Semantic Field is able to offer higher MAP and lower MNL compared to Keyword. The improvement is consistent across all the six categories of concepts. The results basically indicate that Semantic Field is able to include more positive samples and rank them higher than Keyword. A typical example is the concept "animal", where Semantic Field can recall most positive samples although the samples are not tagged with "animal". Another example is "whales", Semantic Field can assign a higher score for the images with tags "whales", "ocean" or "sea" which are more convincing to concept "whales", thus the MAP is improved from 0.5850 to 0.7080.

Table 4: Performance of concept detection by replacing 0%, 25%, 50% and 75% of the ground-truth positive samples with noisy ones.

Categories	MAP			
	0%	25%	50%	75%
people	0.2090	0.1740	0.1430	0.0940
objects	0.1960	0.1720	0.1310	0.0710
scene/location	0.2840	0.2620	0.2170	0.1450
event/activities	0.1100	0.0900	0.0650	0.0360
program	0.1350	0.1080	0.0710	0.0140
graphics	0.1800	0.1130	0.1280	0.0070
all concepts	0.2220	0.1980	0.1590	0.0970

To further investigate the effectiveness of our proposed method, we conduct another experiment by replacing a part of the ground-truth positive samples with randomly chosen negative samples. Three new positive training sets are therefore generated, and each set has respectively 25%, 50% and 75% of false samples. In other words, the MNL (mean noise level) of these new pseudo training sets are 0.25, 0.5 and 0.75 respectively. Table 4 shows the simulation results. As expected, the MAP performance degrades when the percentage of noisy samples increases. For comparison, as shown in Table 2, our Semantic Field method produces a MAP of 0.1660 with a noisy level of 0.5360 (about 54% noisy samples) in terms of MNL, while here when only 50% of noisy samples are used, the performance drops to 0.1590. This again shows the advantage of Semantic Field, which is able to pick up contextually relevant images. As shown in Figure 5, the contextually relevant images, though not completely noise-free, are helpful for concept detection in many cases.

4.3 Detection with Fixed Number of Positive Samples

In practice, the number of positive samples is very hard to predict. In addition, negative training examples also need to be carefully sampled. In this experiment, we adopt one of the most common approaches in sampling pseudo positive and negative samples. The setting is as following. For each target concept, the 161,789 training images are ranked according to their relevancy. A fixed number of 1,000 top-ranked images are then picked as pseudo positive samples, while another 5,000 images are collected as pseudo negative examples from the bottom of the list.

Table 5: Concept detection performance with fixed top-1000 ranked images as pseudo-positive training samples. Note that K indicates the number of nearest neighbors considered by Voting.

Categories	MAP			
	Semantic Field	Keyword	Voting (K=1000)	Voting (K=2000)
people	**0.1428**	0.1317	0.1046	0.1082
objects	**0.1480**	0.1052	0.0806	0.0781
scene/location	**0.2035**	0.1607	0.1566	0.1527
event/activities	**0.0280**	0.0100	0.0190	0.0177
program	**0.0550**	0.0250	0.0290	0.0225
graphics	**0.0255**	0.0007	0.0290	0.0220
all concepts	**0.1543**	0.1163	0.1048	0.1021

We compare Semantic Field to two different approaches: Keyword and tag voting (Voting) [16]. Similar to the previous sub-section, Keyword prepares training samples by collecting web images which are tagged with the target concepts. The images are sorted according to the rank list given by Flickr's search engine. Positive and negative training examples are then sampled respectively from the list for learning SVM classifiers. Voting is a simple yet effective scheme proposed in [16] for assessing the relevancy of a tag to an image. The tag relevancy is determined by accumulating the neighbor votes received from visually similar training images. For instance, a testing image receives 4 votes for tags "bridge" if four of its nearest neighbors are tagged with "bridge". Given a target concept, Voting ranks the testing images according to the neighbor votes received from the 161,789 tagged training images. In our implementation, two global features, color moment and wavelet texture, are used for searching the visually similar neighbors. We experiment two settings, with 1,000 and 2,000 nearest neighbors eligible for voting respectively. In [16], it is suggested the number of neighbors can be set between 200 and 20,000, and 1,000 shows the best performance in their experiment.

Table 5 shows the experimental results. Semantic Field outperforms Keyword and Voting with large margins consistently across all of the six categories of concepts. Compared to the result with the number of positive samples assumed to be known (see Table 2), the performance of Semantic Field drops by about 8%. Considering the more practical setting adopted here that the number of positive samples per concept is unknown, the performance of Semantic Field is quite appealing. From our analysis, the 8% performance

drop can be attributed to the fact that fixing training size to 1,000 samples is not appropriate for all types of concepts. Among the 81 concepts, there are 42 concept classifiers with performance degradation. These classifiers include 19 concepts with the number of ground-truth positives less than 1,000, and 15 concepts with the number more than 3,000. One example is the concept "surf", which only has 124 positive samples according to the ground-truth annotation. As a result, its AP drops from 0.1970 to 0.0120 when 1,000 training images are sampled. On the other hand, Voting produces the worst performance. From our observation, visual search gives rise to excessive number of noisy neighbors, which greatly limit the performance of neighbor voting.

Table 6: Quality of the top-1000 ranked pseudo-positive samples, measured by both MAP and MNL.

Categories	MAP		MNL	
	Semantic Field	Keyword	Semantic Field	Keyword
people	**0.7090**	0.6520	**0.6350**	0.6650
objects	**0.6510**	0.5820	**0.5710**	0.6070
scene/location	**0.6130**	0.5450	**0.7170**	0.7390
event/activities	**0.3930**	0.3430	**0.6960**	0.7000
program	**0.7700**	0.2790	**0.3620**	0.8140
graphics	**0.1200**	0.0520	**0.9320**	0.9760
all concepts	**0.6040**	0.5300	**0.6490**	0.6810

Table 6 further lists the quality of the top-1000 pseudo training examples sampled by the tested approaches*. Comparing Semantic Field to Keyword, although the MNL from Semantic Field is only about 3% lower, the detection performance is 32% better than Keyword. This again gives clue that Semantic Field samples more useful samples for concept learning. On the other hand, as indicated in Table 6, Semantic Field shows better quality of pseudo training examples in terms of MAP. This indicates that Semantic Field has a better capability in ranking relevant training images. In this paper, we have not considered the ranking order in classifier training. We believe that by considering ranking, for instance assigning higher weights to higher ranked images, the detection performance can be further boosted.

5. CONCLUSION

We have presented our analytical studies on the sampling of web images and the quality of pseudo training samples for classifier learning. One general guideline for boosting detection performance is by minimizing the noise level of pseudo training examples. In addition, when assessing the quality of training samples, one aspect should be taken into account is the ability of collecting more typical samples. Even when the noisy samples are falsely included as positive samples, it does not necessarily mean that the learning effectiveness will be degraded, as long as the noisy samples can provide useful contextual clue to a target concept. In other words, the capability of ranking pseudo training samples according to their typicality is a plus in addition to the ability of sampling positive examples. One important problem deserving in-depth future study is how to decide the right number of training examples to be sampled.

On the other hand, we have also presented our approach in predicting the semantic field of tag lists for effective image sampling. Compared to the baseline keyword-based sampling, Semantic Field shows better capability in generating pseudo training set with lower noise level and higher ranking ability. By using three knowledge sources, the proposed approach shows its great capability in recovering concepts which are not tagged, disambiguating tags of multiple senses, and excluding misleading tags. Currently, our approach considers only the semantic field of a tag list and not the visual aspect of an image. Future work includes the joint consideration of semantic field and visual information for more objective sampling. In addition, the ranking of training images, specifically the significance of images, can be incorporated into classifier learning to alleviate the effect of noisy samples.

*We do not list the MNL and MAP of Voting since its prediction is not by SVM but based on K nearest neighbor.

6. ACKNOWLEDGMENTS

The work was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119709).

7. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *ACM SIGCHI*, 2007.
[2] A. Ulges et al. Learning automatic concept detectors from online video. *Comput. Vis. Image Understand*, 2009.
[3] D. Liu et al. Tag ranking. In *ACM WWW*, 2009.
[4] G.-J. Qi et al. Transductive inference with hierarchical clustering for video annotation. In *ICME*, 2007.
[5] J. Tang et al. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM MM*, 2009.
[6] K. Bischoff et al. Can all tags be used for search. In *ACM CIKM*, 2008.
[7] Q. Tian et al. A new analysis of the value of unlabeled data in semi-supervised learning for image retrieval. In *ICME*, 2004.
[8] S.-F. Chang et al. Columbia University/VIREO-CityU/IRIT TRECVID 2008 high-level feature extraction and interactive video search. In *TRECVID*, 2008.
[9] T.-S. Chua et al. NUS-WIDE: A real-world web image database from national university of singapore. In *CIVR*, 2009.
[10] Y.-G. Jiang et al. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Trans. on Multimedia*, 12(1):42–53, 2010.
[11] Y.-G. Jiang, C.-W Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, 2007.
[12] D. Jurafsky and J. H. Martin. *Speech and language processing.* Prentice-Hall, 2000.
[13] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev. To search or to label. In *ACM MIR*, 2006.
[14] L.-J. Li and L. Fei-Fei. OPTIMOL: automatic object picture collection via incremental model learning. *Int. J. of Computer Vision*, 2009.
[15] X.-R. Li and C. G. M. Snoek. Visual categorization with negative examples for free. In *ACM MM*, 2009.
[16] X.-R. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Trans. on MM*, 11(7):1310–1322, 2009.
[17] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at TRECVID. In *ACM MM*, 2004.
[18] G. Quénot and S. Ayache. TRECVID 2009 collaborative annotation. `http://mrim.imag.fr/tvca/`.
[19] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.
[20] A. T. Setz and C. G. M. Snoek. Can social tagged images aid concept-based video search. In *ICME*, 2009.
[21] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM MM*, 2001.
[22] G. Wang, T.-S. Chua, and M. Zhao. Exploring knowledge of sub-domain in a multi-resoluation bootstrapping framework for concept detection in news. In *ACM MM*, 2008.
[23] R. Yan, A. G. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *ACM MM*, 2003.
[24] R. Yan and M. R. Naphade. Semi-supervised cross feature learning for semantic concept detection in video. In *CVPR*, 2005.