ARISTA - Image Search to Annotation on Billions of Web Photos

Xin-Jing Wang[†], Lei Zhang[†], Ming Liu[‡], Yi Li[‡], Wei-Ying Ma[†]
[†]Microsoft Research Asia, 49 Zhichun Road, Beijing, China
[‡] Microsoft Corporation, One Microsoft Way, Redmond, U.S.A.
{xjwang,leizhang,miliu,yili,wyma}@microsoft.com

Abstract

Though it has cost great research efforts for decades, object recognition is still a challenging problem. Traditional methods based on machine learning or computer vision are still in the stage of tackling hundreds of object categories. In recent years, non-parametric approaches have demonstrated great success, which understand the content of an image by propagating labels of its similar images in a large-scale dataset. However, due to the limited dataset size and imperfect image crawling strategy, previous work can only address a biased small subset of image concepts. Here we introduce the Arista project, which aims to build a practical image annotation engine targeting at popular concepts in the real world. In this project, we are particularly interested in understanding how many image concepts can be addressed by the data-driven annotation approach (coverage) and how good the performance is (precision). This paper reports the first stage of the work. Two billions web images were indexed, and based on simple yet effective near-duplicate detection, the system is capable of automatically generating accurate tags for popular web images having near-duplicates in the database. We found that about 8.1% web images have more than ten near duplicate and the number increases to 28.5% for top images in search results. Further, based on random samples in the latter case, we observed the precision of 57.9% at the point of the highest recall of 28% on ground truth tags.

1. Introduction

The overwhelming amounts of data on the Web have not only inspired many interesting applications [14][13] but also enabled simple yet effective solutions to many hard problems which have plagued researchers for decades [8][15][7][1].

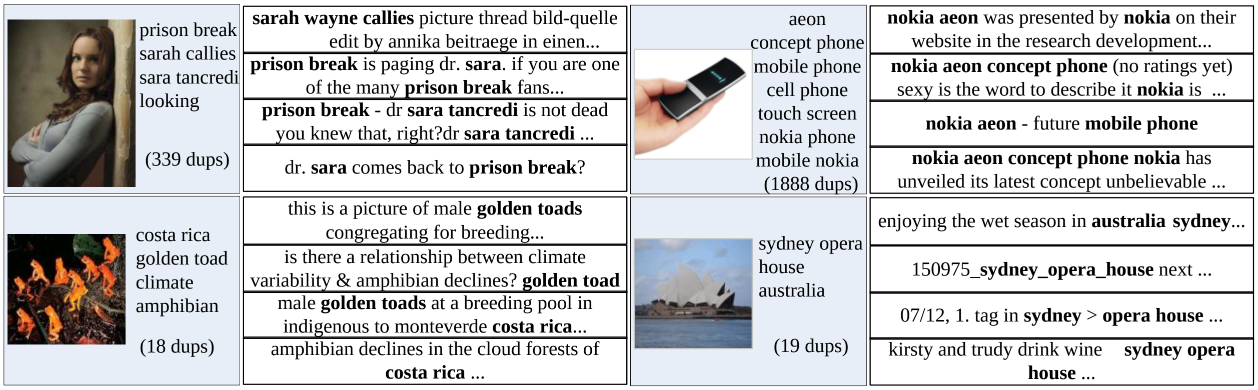
Recently, there is a surge of interest in data-driven approaches for image auto-tagging [15][1][6]. The motivation behind is that if a large enough image database is available, the visually close similar images will possess certain semantic similarity, so their labels can be propagated in between and to tag an unlabeled image. Wang et al. [15] collected 2.4 million high-quality web images with abundant surrounding texts, and tagged a new image by mining common phrases from the surrounding texts of its visually similar images. Torralba et al. [1], furthermore, collected about 80 million tiny images of 32x32 pixels and confirmed that with simple nearest neighbor methods, the recognition performance improves when the image database expands. This is also the largest dataset used in literature. Deng et al. [6], on the other hand, structuralized 3.2 million web images and observed improved recognition performance assisted by image ontology.

Though millions of images have been used, they still occupy a small portion of web images[1] as well as cover a biased small subset of concepts in the real world. For example, the images used by Wang et al. [15] are mostly scene photos, and Torralba et al.[1] and Deng et al. [6] crawled photos tagged by non-abstract nouns in WordNet [3]. Many image concepts which are more closely connected to people’s everyday life were ignored, such as paintings, celebrities, movies, products, and logos.

In this paper, we introduce the Arista project (lARge-scale Image Search To Annotation), as an attempt to investigate the performance of image annotation on a real web-scale image dataset. Two billion web images were leveraged for this investigation. The goal is to build a practical image annotation engine which is able to automatically annotate images of any popular concepts[2]. This paper provides a detailed report of this research effort in its first stage: we investigate the value of near-duplicate images in object recognition on a two billion dataset. The study in this stage shows that such a near-duplicate-based annotation approach can address many difficult concepts, which cannot be effectively represented by existing visual features, and are also beyond the scope of WordNet [3]; therefore, they are much more difficult concepts than those handled by traditional object recognition approaches [10]. Examples of such concepts are celebrity, logo, product, landmark, poster, pattern, etc., which are popular concepts and are more likely to be duplicated. On the other hand, near-duplicate images are of a special type of visual-

[1] Flickr hosts >4 billion photos, and Google maintains more.

[2] We define a “popular concept” as a word or a phrase that has at least ten highly relevant images on the Web.


prison break
sarah callies
sara tancredi
looking
(339 dups)

sarah wayne callies picture thread bild-quelle edit by annika beitraege in einen...

prison break is paging dr. **sara**. if you are one of the many **prison break** fans...

prison break - dr **sara tancredi** is not dead you knew that, right?dr **sara tancredi** ...

dr. **sara** comes back to **prison break**?

aeon
concept phone
mobile phone
cell phone
touch screen
nokia phone
mobile nokia
(1888 dups)

nokia aeon was presented by **nokia** on their website in the research development...

nokia aeon concept phone (no ratings yet) sexy is the word to describe it **nokia** is ...

nokia aeon - future **mobile phone**

nokia aeon concept phone nokia has unveiled its latest concept unbelievable ...

costa rica
golden toad
climate
amphibian
(18 dups)

this is a picture of male **golden toads** congregating for breeding...

is there a relationship between climate variability & amphibian declines? **golden toad**

male **golden toads** at a breeding pool in indigenous to monteverde **costa rica**...

amphibian declines in the cloud forests of **costa rica** ...

sydney opera
house
australia
(19 dups)

enjoying the wet season in **australia sydney**...

150975_**sydney_opera_house** next ...

07/12, 1. tag in **sydney** > **opera house** ...

kirsty and trudy drink wine **sydney opera house** ...

Figure 1. Examples showing that surrounding texts of near-duplicates have common terms which hit the semantics of a query image. The tags inside the image blocks are our annotation outputs. The common terms of each near-duplicate are highlighted in bold. Note that the detected tags are very specific. This is in contrast to most existing works that tend to generate general terms like sky, city, etc.

ly similar images, and near-duplicate detection is a well-defined problem as contrast to visual similarity search. In this work, we address those concepts[3] that favor the near-duplicate-based search-to-annotation approach and leave the rest to more sophisticated object recognition techniques as our future work.

It is worth highlighting that near-duplicate images were generally regarded as annoying garbage which wastes the resource of search engines and lowers the diversity of search results [8]. To our knowledge, this is the first work to demonstrate the value of near-duplicate images.

In this study, we investigate the percentage and concept distribution of web images that have near duplicates, evaluate the effectiveness of the proposed near-duplicate-based annotation approach, and study the impact of dataset size as well as the number of near duplicate images on the annotation performance. Base on 70k randomly selected web image from the 2 billion dataset, we found that about 22% have near-duplicates, and about 8.1% have more than ten. The latter number increases to 28.5% for top images in search results, and based on random samples from these queries, we obtained the tagging precision of 57.9% when the recall peaks at 28% according to manually labeled ground truths.

The paper contains five parts. In Section 2, we present our insights and expectations, and then report in detail the statistical analysis and observations in Section 3 and Section 4. A few potential applications are discussed in Section 5 and Section 6 outlines the conclusions.

2. What we can expect from a large image DB

The key hinder factor of computer vision research such as object recognition, scene recognition, and image search, is the semantic gap between existing low-level visual features and high-level semantic concepts. Previous machine learning and computer vision approaches [9][10] attempted to directly map visual features to textual keywords. Since these two types of features are heterogeneous, the intrinsic mapping function is totally in the dark.

On the contrary, the data-driven approaches [15][1] leverage a search-to-annotation strategy and an intermediate image set (a group of visually close similar images) to build the connections. Since the similar images are partly labeled, they provide ground truth knowledge of the mapping between images and textual keywords. Therefore, the annotation problem in such approaches is simplified as first to map between visual features, i.e. to measure the visual similarity of a query image against the intermediate dataset, and then to determine which keywords should be propagated to the query image. In this way, it greatly reduces the difficulty.

Now the key problem is how to ensure visual similarity to represent semantic similarity. Torralba et al. [1] observed that when the dataset grows, the probability of finding a visually close similar image increases sharply, and when the visual similarity exceeds a certain threshold, the probability of the corresponding images belonging to the same visual class grows rapidly, which means visual similarity is approaching semantic similarity. These suggest that a large dataset is important - with a very large dataset, we can set a strict threshold on visual similarity to ensure semantic similarity, and at the same time have a large chance of finding enough number of close similar images.

In short, assisted by a large-scale (partly) labeled visually close similar image dataset, the semantic gap can be greatly passed by. This is the key idea of the data-driven image annotation approaches [15][1][6].

We are curious about that: 1) does the number of accurate tags increase along with the growing dataset size, and 2) what is the enough scale of the dataset so that even if it further expands, the annotation precision will not increase? We try to answer these questions in this study.

In contrast to those research efforts which attempted to detect and remove duplicate web images [2][8], we treat them as a valuable population among web images. Gener-

[3] These concepts make up of a much larger vocabulary beyond those from the existing works [15][1][6][13].

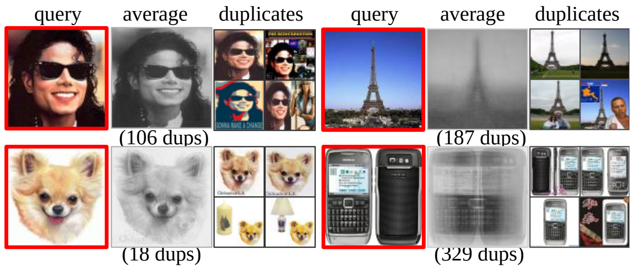


Figure 2. Examples of near duplicates. Query images are highlighted in red blocks. From left to right: query image, average image of near duplicates, and examples of near duplicates.

ally, when an image is duplicated in multiple webpages, the author of a webpage will attach his/her own words to the near duplicate. Though different surrounding texts are associated with each near duplicate, the most important (and correct) terms to an image are very likely to be repeated among the majority of its near duplicates, and thus can be detected to annotate this image. Figure 1 shows a few examples. The tags inside each query image block are the outputs; to their right, each row shows some surrounding texts of one near duplicate.

From Figure 1 we can see that the query images are tagged with very specific and accurate terms, e.g. celebrity names, product models, etc. This verifies the intuition that specific (and semantic) terms appear more frequently among near duplicates, so that there is a larger chance for the images to agree on them. Contrarily, due to the lack of description power of visual features, especially those such as the global features used in [15] and pixel-level features used in [1], if the dataset is not large enough, the images detected which are visually similar may be far from semantically similar. As a result these images tend to agree on general terms. For example, for the last query image in Figure 1, if the dataset is too small to cover enough images of Sydney Opera House with blue sky, then using the features of [15] or [1], images with dominant blue color will be retrieved as visually similar ones. Therefore, the query will tend to be tagged by “sky” and “beach” rather than “Sydney Opera House”. We will explore in details the impact of dataset size on tag accuracy in Section 4.3.

3. Statistics on near-duplicate images

The first necessary understanding of the effectiveness of searching near duplicates to annotate a query image is 1) how many images have near-duplicates? Or say, how large is the population of images that can be accurately annotated with the proposed method? and 2) concretely what concepts tend to have near duplicates?

We address these questions in this section.

3.1. The near-duplicate image detection approach

The problem of near-duplicate detection has been studied for years [2][8], however, how to efficiently detect as many as possible near duplicates on billions of images is still very challenging.

Table 1. Duplicate statistics on 70k random images from 2B DB

#dups	=0	1~9	10~99	100~500	>500
%	78.03	13.86	5.30	1.48	1.32

We allocated several hundreds of machines and dumped two billion web images from a large web image database. Based on them, we set up an evaluation platform and built an index to perform near duplicate search.

Due to confidential reasons, we cannot provide the details. However, theoretically any near duplication detection techniques can be applied. For example, Zhang et al. [11] used a PCA-based hash function to map a high-dimensional feature to a 32-bits hash code. They first extracted visual features on $k \times k$ regular grids of an image, and then transformed these features to a pre-learnt low-dimensional PCA space. The reported computational complexity to find all near duplicates among n images is about $O(nlog(n))$. If a hash table is used to detect hash collision, the computational complexity can be reduced to a constant level which is far less than n. A similar approach was adopted by Wang et al. [15].

We are able to detect one near duplicate in 70 milliseconds on average (evaluated on about 200k image queries). Figure 2 discloses the effectiveness. The four randomly selected query images are marked in red blocks. To their right are the corresponding average images and a few examples of their near duplicates. Intuitively, the average image of a set with many false alarms will be much more blurred than that of a set with few diverse images. Though many largely modified near duplicates are available, the average images are still recognizable. This illustrates the effectiveness of our near duplicate detection technique.

3.2. Duplicate statistics on random image queries

To evaluate how large is the image population having near duplicates, we randomly collected about 70k web image from the 2 billion dataset. The duplicate statistics is shown in Table 1. It shows that about 22% web images have near duplicates. This table provides us a rough idea of the percentage of images which can be successfully tagged by the near-duplicate search-to-annotation approach. Assume that it requires at least ten near duplicates to effectively tag one query image; it suggests this approach is able to tag 8.1% images on the Web.

3.3. The query sets on image concepts

To evaluate how our approach addresses the image concepts favored by web users, we collected two query sets from the image query log in July, 2009: one is the top query set to represent popular concepts submitted by web users and the other is the random query set. In addition,

Table 2. Statistics of the textual query sets used in the study

	top	random	animal	landmark
#text query	3,369	4,785	153	32
avg. freq	7951.7	4,871.9	-	-
#img query	65,274	67,678	3,011	596

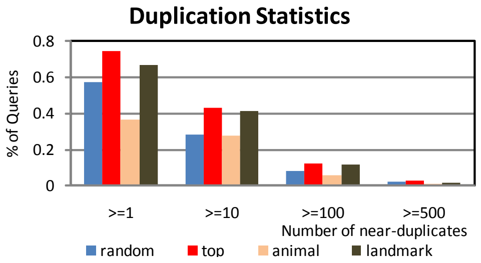


Figure 3. The query distribution against the number of near-duplicates follows the power law.

we used another two collected from public websites[4], which contain the concepts generally considered by traditional object recognition approaches [10][1][6].

Table 2 shows the statistics on the textual queries and their average query frequencies. We submitted each textual query to a commercial image search engine and collected the top twenty images[5] for the analyses in Section 3.4 and Section 3.5; the sizes of resulted image query sets are listed in the last row.

3.4. Duplicate statistics on image concepts

The distribution of the percentage of image queries against the number of near duplicates follows the power law. As shown in Figure 3, about 36.5% animal images and 67.0% landmark images are duplicated at least once. Most of the images have less than 500 near duplicates, while the very "hot" images which have more than 500 duplicates occupy relatively a small population.

Moreover, images of top query concepts are more likely to have duplicates. About 74.4% top image queries have at least one near duplicate, which is about 17.2% higher than the random set, and the percentage of top queries which have more than ten near duplicates largely surpasses that of the random set.

Recall that about 8.1% web images have at least ten near duplicates. As for the top images in search results, this number increases to 28.5% on random queries.

3.5. Category statistics on concepts having dups

To understand which image concepts are more likely to be addressed by this approach, we categorized the random queries by mapping them against the ODP ontology [12]. The reason that we prefer ODP to WordNet [3] is that ODP is better aligned with Web search, which have a much larger coverage on the web queries than WordNet. For example, WordNet does not cover terms like "Michael Jackson" or "xbox 360".

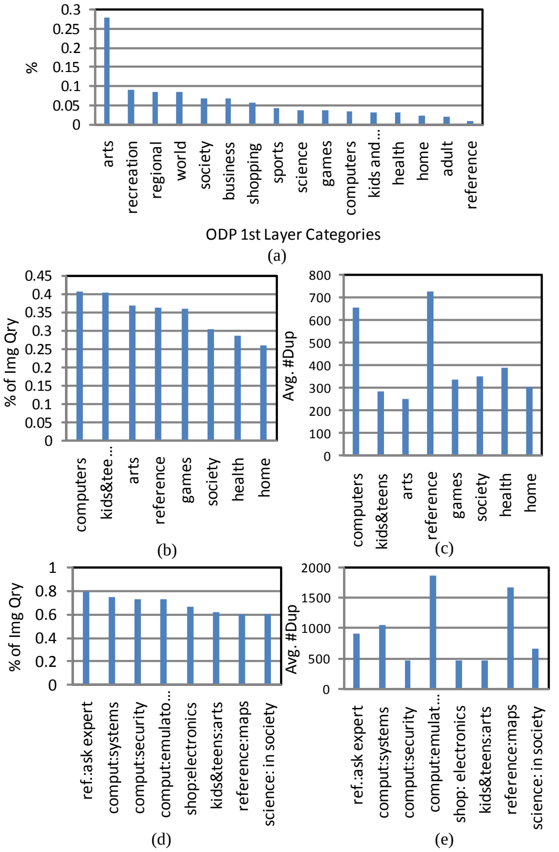


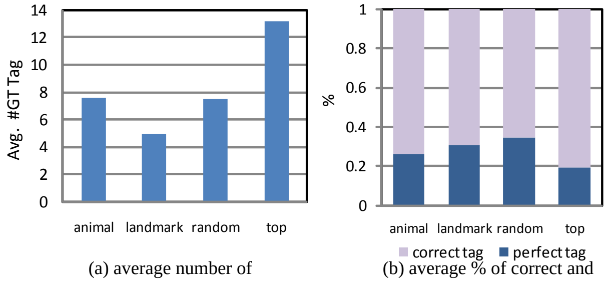
Figure 4. Query concept distribution against ODP of the "random" query set. (a) Percent of seed queries covered by the 15 ODP root categories. (b) The top 8 ODP root categories having the largest image query populations which have >10 duplicates. (c) Average number of duplicates corresponding to (b). (d) The top 8 ODP second-tier categories on image queries having >10 duplicate. (e) The corresponding average number of (d).

The query-ODP mapping is like this: a seed query term is first submitted to ODP [12], and the top returned ODP tree path is assumed as its ODP category. Each image query indexed by the term is then assigned to the category.

Figure 4 summarizes the concepts that most probably have near duplicates. Figure 4(a) shows the distribution of the seed query terms on the ODP root categories. It shows that most user queries belong to the "arts" category, e.g. celebrities, movies, music. Figure 4(b) illustrates the top eight ODP root categories which have the largest query populations; only the queries having more than ten near duplicates were considered. Figure 4(a) and Figure 4(b) together suggests that though "computers" is not the most

[4] animal: http://en.wikipedia.org/wiki/List_of_animal_names. landmark: http://www.nationsonline.org/oneworld/most_famous_landmarks.htm.

[5] Considering that web users generally browse top search result before they leave, the relevance of top search results is more important to search engines.



(a) average number of ground-truth tags

(b) average % of correct and perfect tags in ground-truth

Figure 5. Statistics of the labeled ground-truth.

frequent query type, images of this category tend to be duplicated. Examples are "intel", "cool wallpapers", "c3d". Moreover, "arts" is not only the most frequent query type, but also its images tend to be highly duplicated. Figure 4(d) and Figure 4(e) show respectively the top eight second-tier ODP categories of highly-duplicated concepts and their average duplication numbers. Example are "question mark", "blackberry storm", "smile icon", "120 GB", "baby polar bear". Most of such queries are graphics, icons, computers and consumer electronics.

4. Statistics on near dup-based annotation

This section presents our detailed investigations on the effects of dataset size and duplication number on annotation performance. The text descriptions of an image are made up of the image caption, URL, title of the landing page, and surrounding texts in the landing page.

4.1. Experimental settings

For each query set, we randomly selected 800 query image, each of which has no less than 3 duplications. Seventeen datasets scaling from 2.4 million to 2 billion were randomly sampled from the 2 billion dataset. Note that 2.4 million is the dataset size used by Wang et al. [15].

We independently performed the auto-tagging approach on each dataset and aggregated the annotations of the same image to be judged by ten human subjects, which gave the ground truth tags. Bing and Google image search results were provided to assist the labelers in order to ensure the quality of their labels. We did not use existing benchmark datasets [4][6] not only because most of the image concepts suitable for this study are not included in those datasets (e.g. celebrities, products), but also because the tags they provided are inadequate for practical web image annotation. For example, ImageNet [6] was collected against WordNet [3]; it cannot suggest tags such as "impressionism" to Van Gogh's masterpiece "Sunflower".

We also required the labelers to mark a true positive tag as "perfect" or "correct". "Perfect" means a tag is not only accurate but also specific, while "correct" means a tag denotes general concepts of a query image. For example, for a Youtube logo, a perfect tag is "Youtube", while correct tags are "video sharing", "social networking", and for a pigeon image, a perfect tag is "pigeon", while correct tags include "bird", "animal", "feather" etc. Figure 5 shows some statistics on the ground truth labels.

4.2. Tag mining Methods

4.2.1 SRC [5]

Wang et al.[15] showed the effectiveness of mining tags with the Search Result Clustering (SRC) [5] technique, which is based on a pre-learnt regression model to score n-grams in search results. Here we evaluate this method.

4.2.2 Majority voting (MV)

We are also interested in whether a simple voting method works in this scenario and how good it is. We tokenize the text descriptions of a query as well as its near duplicates into terms and remove the stopwords. Each term is then scored by its document frequency (DF), which is the number of near duplicates containing this term. Terms whose $DF \leq 2$ are removed; this is the tradeoff between annotation precision and coverage on query (i.e, the population of query images that actually get tagged).

It is worth highlighting that in order to avoid unnecessary parsing errors, no parsing tools were used. Therefore our MV produces no phrases. This is contrary to SRC [5], which possesses a key advantage of producing phrases.

To compensate possible performance degradation, we propagate the score of a ground-truth phrase to a term once the phrase contains this term, which gives the *upper-bound* performance of the majority voting tagging method.

4.3. The effect of dataset size

4.3.1 Tagging performance vs. dataset size

Figure 6 shows how the tagging average precision (AP) and average recall (AR) varies against the increased dataset size. The real curves denote using the SRC [5] method, while dotted curves represent the best performance that MV achieved. A few observations are:

1) When dataset size increases, so does average recall, no matter which tag mining method is used. Intuitively, larger dataset implies more near duplicates, which enlarges the probability of discovering new accurate terms.

2) When dataset size is larger than 300 million, average precision of the SRC method converges, while that of MV is dropping. Intuitively, larger dataset implies larger candidate term set and more noisy terms. Therefore, it indicates that SRC is more robust than MV in handling noise. Moreover, SRC surpasses MV on these dataset sizes.

3) On the landmark dataset, SRC seems inferior to MV. The reason is that the landmark names are generally phrases, such as "Taj Mahal", "Forbidden City", "Schonbrunn Palace". MV is actually assigned much more credit in this case because we assume an MV tag to be correct as long as it matches a part of a ground truth phrase.

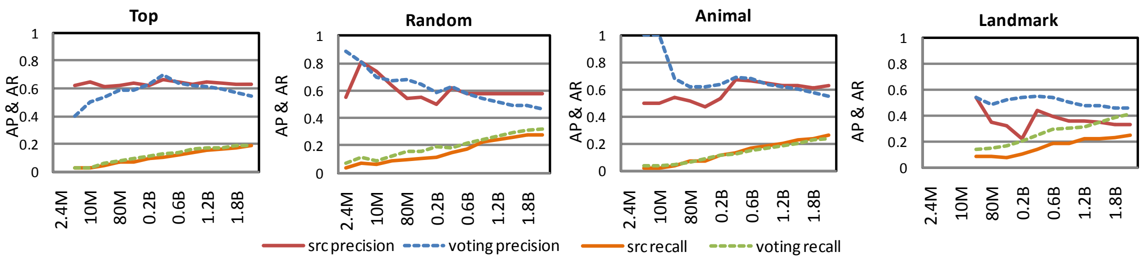


Figure 6. The effects of dataset size (2.4million ~ 2 billion). Solid lines show the average precision/recall of using SRC, and dotted lines are of MV. The AP performance converges when the DB size increased to 300 million, while average recall is still improving.

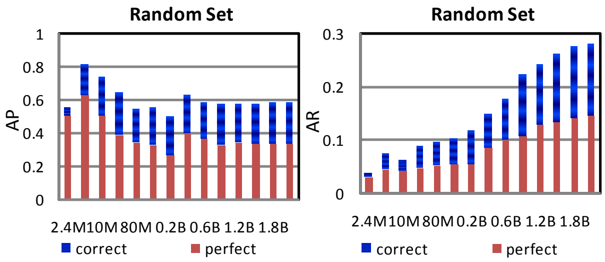


Figure 7. Specific effects of dataset size on the perfect and correct tags. Evaluated on the random query set using SRC.

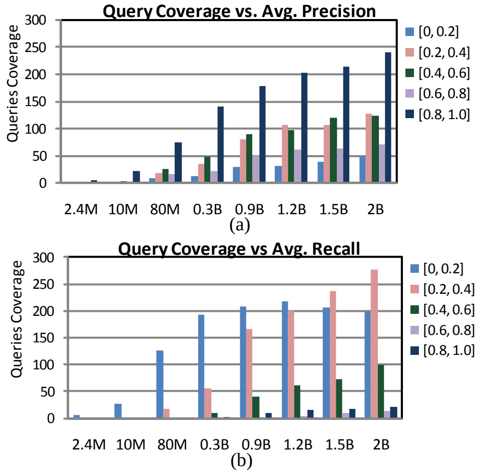
4) Irregular AP curve segments were observed when the dataset size is smaller than 200 million, especially for SRC. The reason is that a query image may not always be tagged in each dataset. Therefore, the number of images that truly attends the evaluation is varying. Using the random set as an example, when the dataset size is under 200 million, only tens of queries were tagged, which is not enough to provide a trustworthy statistical measure on the performance. Such a problem of low coverage on query images is even severe on the animal and landmark sets --- when the dataset size is smaller than 5 million, no animal queries were tagged, and when it is under 30 million, no landmark queries were tagged.

Figure 7 specifically illustrates the effect of dataset size on perfect and correct tags. It suggests that:

1) The AP performance of SRC on correct tags is not dropping but climbing as dataset size grows, which indicates that it can always detect new accurate terms.

2) Both the AR performance on perfect and correct tags improves as the dataset size increases. Specifically, the chance of detecting perfect tags on a 2B dataset is about three times of that on a 2.4M dataset, and is about five times when detecting correct tags.

Discussions. Torralba et al. [1] witnessed improved AP when the dataset size increases, which seems conflict with our observations of converged AP performance. In fact, this is resulted from the different experimental settings. Torralba et al. [1] formulated the problem as a classification task; they fixed the number of object classes and measured AP against this fixed ground truth. However in our evaluation, when the dataset grows, so does the average number of tags detected. The fact that the AP curve converges while the AR curve is still climbing exactly proves that the growing dataset size really brings more useful terms.



(a)

(b)

Figure 8. The number of queries being tagged and the corresponding AP and AR score distributions on the datasets, evaluated on the random set. The larger the dataset, the more queries will be annotated.

Figure 9 gives a few examples of the query images and their detected tags along with the increased dataset size. It shows that larger dataset size ensures more accurate tags.

4.3.2 *Tagging performance vs. coverage of queries*

Intuitively, when the dataset grows, the chance of a query getting tagged will increase along with the expanded near-duplicates set. Figure 8 proves this intuition. Figure 8(a) and Figure 8(b) shows respectively the coverage of queries on different datasets against the bins of average precision and average recall. A few interesting conclusions can be drawn from this figure:

1) See Figure 8(a), both query coverage and AP performance improve as dataset size grows.

2) See Figure 8(b), when dataset is larger than 1 billion, the query set which has a small AR score between 0 and 0.2 shrinks gradually. Contrarily in the rest cases, when the dataset grows, the number of queries which have high AP and AR scores increases. This again demonstrates that

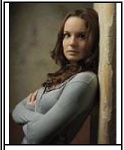
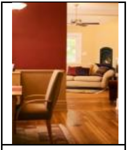
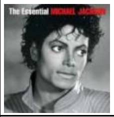
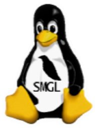
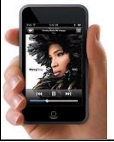
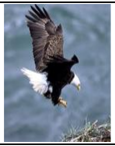
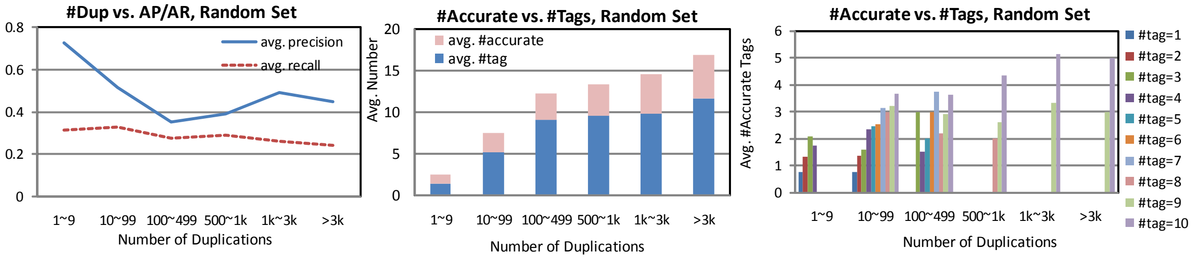
	2.4 M	80M	2B		2.4M	80M	2B
	(no results)	(no results)	*prison break*, **sarah callies**, **sara tancredi**, looking		(no results)	house paint, color	*house*, *paint*, wantatoos, house painting, hardwood floor, interior design
	**michael jackson**	**michael jackson**, *rock pop*	**michael jackson**, sony music, *cd dvd*, *entertainment music*, *pop rock*		linu, *logo*	server, *software*, *logo*, credit card processing, *operating system*	**penguin**, *open source*, *virtual server*, *logo*, *operating system*
	**ipod touch**	**apple ipod**, *mp3 player*, **iphone**, wi fi, touch screen	**apple ipod**, *mp3 player*, wi fi, media player, **touch screen**, **mobile phone**		(no results)	(no results)	**bald eagle**, **haliaeetus leucocephalus**, endangered species, fish wildlife, *eagle flight*

Figure 9. Annotation examples vs. dataset size. Bold-faced tags are perfect terms labeled by human subjects and italic ones are correct terms. Due to space limit, only the top five tags are shown. This figure suggests that larger dataset size ensures more accurate tags.



(a) (b) (c)

Figure 10. The effect of duplication number on annotation performance. (a) Curves of AP and AR scores on the query sets of different duplication numbers. (b) The average number of tags and correct tags detected on the query sets. (c) Distributions of the average numbers of positive tags among the top ten detected tags on different query sets.

the larger the dataset, the better chances of a query to be accurately and adequately tagged.

3) 80 million is still far from enough either for good recognition performance or for the coverage of image concepts, though Torralba et al. [1] have shown that it provides reasonable nearest neighborhoods of images.

4) The number of queries which have high performance (i.e. precision/recall in between 0.8 and 1.0) tends to converge when dataset size exceeds one billion.

4.4. The effect of duplication number

Figure 10 illustrates the effect of the duplication number of a query on its annotation performance.

From Figure 10(a), it seems that having more duplicate images does not necessarily mean better annotation performance. Figure 10(b) investigates the problem. It can be seen that the number of tags grows rapidly when the duplication number increases from several to hundreds, and slows down after the duplication number exceeds a hundred. Though the number of accurate tags detected is also growing, its pace is slower. Possibly this is caused by two factors. One is that the tag mining approach is effective. The other is that the stopword list we used is not abundant. Many images were tagged with terms such as “myspace”, “facebook”, “asp”, which are typical noisy terms in the Web scenario. However, whether such words can be regarded as stopwords is an open question. For example, in the case of a facebook logo, “facebook” will be an accurate tag, but it is not as for an image shared in facebook. Figure 10(c) zooms into Figure 10(b) and shows the correlation between the number of accurate tags and detected tags against the duplication number. In the case of few near duplicates (i.e. 1-9), all the queries have no more than four tags detected. Contrarily, when the number exceeds three thousands, all the queries have more than nine tags, and the more near duplicates, the more detected tags. Interestingly, the queries which have duplication number between 10 and 99 cover all the cases of tag number.

5. Interesting applications

A key advantage of search-based annotation is that it tags an image with an open vocabulary. Ideally, as long as we can find enough partly labeled similar images correlated with a new concept, we can recognize a new image with this concept. The approach studied in this paper specifically addresses popular concepts such as celebrities,

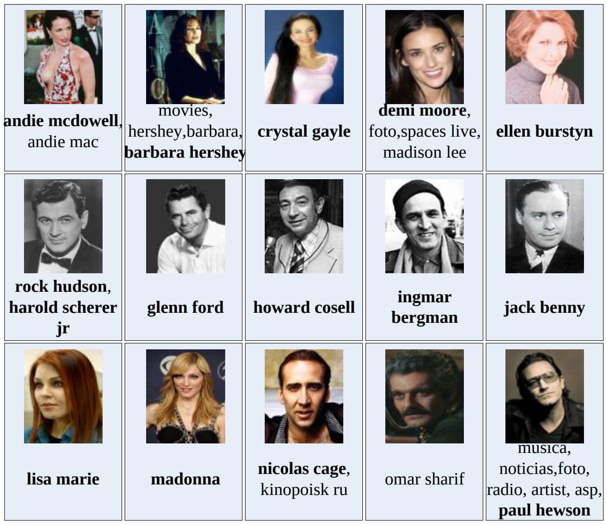

andie mcdowell, andie mac
movies, hershey,barbara, **barbara hershey**
crystal gayle
demi moore, foto,spaces live, madison lee
ellen burstyn
rock hudson, **harold scherer jr**
glenn ford
howard cosell
ingmar bergman
jack benny
lisa marie
madonna
nicolas cage, kinopoisk ru
omar sharif
musica, noticias,foto, radio, artist, asp, **paul hewson**

Figure 11. Face book of celebrities. Tags are the outputs on the 2B dataset using SRC method. Bold-faced tags are perfect labels.

consuming electronics, landmarks, and entertainment images. It provides great potential to develop interesting and useful applications. We discuss a few here and leave detailed investigations to the future work.

5.1. Generating a large face book

An interesting application is to construct a large face book of celebrities. Figure 11 shows a few results on celebrity queries. The bold-faced terms suggest the ground truth celebrity names. Though there are noisy tags, it can be seen that the approach successfully identified all the celebrities' names.

We believe a promising solution of person identification or face recognition can be developed based on this approach. By applying some name identity extraction or term categorization techniques to the tags, and utilizing face detection result, we are able to build a very large face book of celebrities.

5.2. Building a visual dictionary for landmarks

Though it is highly demanded, existing location dictionaries of good coverage and quality are generally edited manually. As to our knowledge, there are no complete visual dictionaries on famous landmarks in the world. Intuitively, we can adopt the similar idea in Section 5.1 to build a large visual dictionary for landmarks. We will investigate the feasibility in our future work.

6. Conclusion and future work

Search-to-annotation methods have received significant attention in recent years in addressing the challenging problem of image understanding. Rather than resorting to complicated machine learning or computer vision techniques, such a kind of approach leverages a huge amount of partly labeled web images to bypass the semantic gap.

Previous approaches in this direction generally play with a few millions of images, which still occupy a small population of web images and cover a biased subset of the world's image concepts. In this work, we tested the same idea on a real web-scale dataset with billions of images. This paper reports our discoveries in the first stage, i.e. annotating a new image by mining key phrases from its near duplicates. We found that when the dataset size increases, more accurate tags are discovered. Though tagging precision becomes constant on a scale of hundreds of millions of images, recall keeps improved. Moreover, a larger dataset means larger coverage on query images.

In the future, we will move on to investigate the performance of using visually close similar images rather than near duplicates, and test the effectiveness of existing visual features. This will inevitably bring in new index challenges to the annotation system. We will study the large scale indexing problem as well along this direction.

References

[1] A. Torralba, R. Fergus, and W. Freeman. 80 Million tiny images: a large dataset for non-parametric object and scene recognition. IEEE T-PAMI, 30(11):1958–1970, 2008.

[2] B. Wang, Z. Li, M. Li, and W.Y. Ma. Large-scale duplicate detection for web image search. In ICME, 2006

[3] C. Fellbaum. WordNet: An Electronic Lexical Database. Bradford Books, 1998.

[4] G. Griffin, AD. Holub, P. Perona, The Caltech-256, Caltech Technical Report.

[5] H.-J. Zeng, Q. He, Z. Chen, and et al. Learning to cluster web search results. SIGIR. 2004.

[6] J. Deng, W. Dong, R. Socher, and et al. ImageNet: a Large-scale hierarchical image database. CVPR, 2009.

[7] J. Hays and A. A. Efros. Scene completion using millions of photographs. SIGGRPAH, 2007.

[8] J. J. Foo, J. Zobel, R. Sinha, and S. M. M. Tahaghoghi. Detection of near-duplicate images for web search. In CIVR, 2007.

[9] J. Li, J.Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach. In IEEE T-PAMI, 25(9):1075-1088, 2003.

[10] K. Barnard, P. Duygulu, N. de Freitas, and et al. Matching words and pictures. JMLR, 3:1107–1135, 2003.

[11] L. Zhang, L. Chen, F. Jing and et al. EnjoyPhoto: a vertical image search engine for enjoying high-quality photos. ACM Multimedia. 2006.

[12] ODP. The Open Directory Project. http://dmoz.org/. 2009.

[13] R. Datta, D. Joshi, J. Li, J. Z. Wang. Image retrieval: ideas, influences, and trends of the new age, ACM Computing Surveys, 40(2):1-60, 2008.

[14] T. Yeh, K. Tollmar, and T. Darrell. Searching the Web with mobile images for location recognition. CVPR, 2004.

[15] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. AnnoSearch: image auto-annotation by search. CVPR, 1483-1490. 2006.

[16] Y. Ke and R. Sukthankar. PCA-sift: A more distinctive representation for local image descriptors. CVPR, 2004.