Reliable Tags Using Image Similarity

Mining Specificity and Expertise from Large-Scale Multimedia Databases

Lyndon Kennedy
Yahoo! Research
Santa Clara, CA
lyndonk@yahoo-inc.com

Malcolm Slaney
Yahoo! Research
Santa Clara, CA
malcolm@ieee.org

Kilian Weinberger
Yahoo! Research
Santa Clara, CA
kilian@yahoo-inc.com

ABSTRACT

This paper describes an approach for finding image descriptors or tags that are highly reliable and specific. Reliable, in this work, means that the tags are related to the image's visual content, which we verify by finding two or more real people who agree that the tag is applicable. Our work differs from prior work by mining the photographer's (or web master's) original words and seeking inter-subject agreement for images that we judge to be highly similar. By using the photographer's words we gain specificity since the photographer knows that the image represents something specific, such as the Augsburg Cathedral; whereas random people from the web playing a labeling game might not have this knowledge. We describe our approach and demonstrate that we identify reliable tags with greater specificity than human annotators.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Human Factors

Keywords

tagging, reliability, specificity, image similarity

1. INTRODUCTION

As the volume of multimedia available on the web grows, so does the need for tools for searching, organizing, and exploring this immense resource. A variety of research efforts are currently showing great promise in a wide array of applications over this media, such as multimedia retrieval, automatic image and video annotation, and multimedia summarization.

A common aspect of many of these new systems for processing multimedia is that they rely heavily on the ability to solicit reliable annotations for a large number of example images. Methods for learning automatic classifiers for various visual objects rely on large sets of labeled images. Visual image search can be greatly improved if images are assigned reliable tags. Annotation presents a number of difficulties, however. The provided labels are often noisy and unreliable. Furthermore annotators may only have a cursory knowledge of the visual objects in the images and may not be able to provide any specific details.

A number of approaches have been proposed for increasing the reliability of image labels, ranging from utilizing the input from multiple annotators in a game scenario [10] to leveraging the tags of visually similar images to re-order or re-weight the labels associated with a given image [6, 7]. In this work, we propose a new framework for gathering reliable image labels that lies somewhere between these two approaches. We leverage a large database of tagged photographs and discover pairs of visually similar images. We

australia, **uluru**, sunrise, honeymoon, sacred, **ayers**, rock, katatjuta, nature, clouds, outback, brush, orange, shrubs, desert, dry, thruhike98

uluru, sunrise, climb, beautiful, rock, **ayers**, chain, Ayers Rock, **Australia**, geotagged

(a) Photos of Uluru (Ayers Rock) in Australia

ZenpukujiPark, Ogikubo, **CommonKingfisher**, カワセミ

birdwatching, Guarda-Rios, wildbird, **CommonKingfisher**

(b) Photos of Common Kingfishers

Figure 1: Examples of visually similar images with highly specific shared tags found automatically with our proposed system. The specific tags are shown in bold.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WSMC'09, October 23, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-761-5/09/10 ...$10.00.

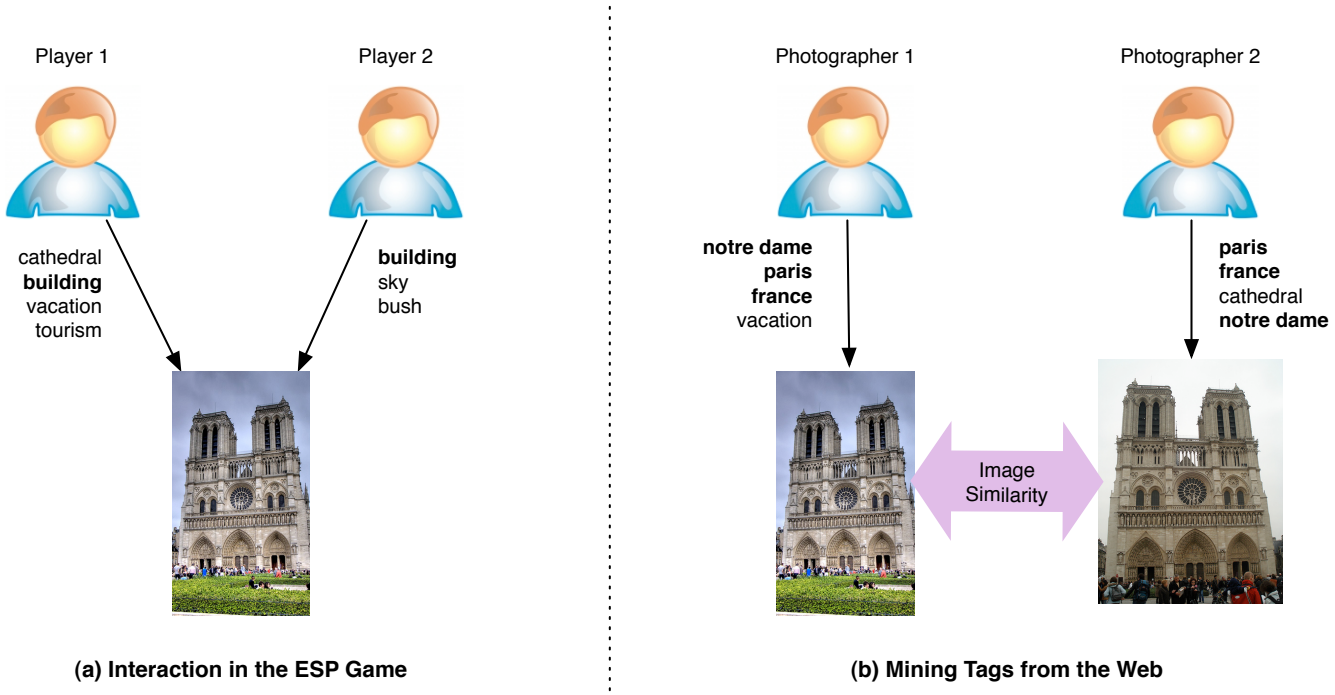

Figure 2: Tagging activity and interaction in (a) the ESP game and (b) our proposed web mining approach.

then make the assumption that, if there are one or more matching tags entered by separate people on each of these visually similar images, then those tags are likely to be related to the image content, and therefore reliable and useful for applications like training visual models or search.

We conduct a study using a database of over 19 million images collected from a photo sharing website. We find that the method can significantly increase the relevance of the tags associated with the images. We further find that photographers (the image creators) are more adept than random annotators at providing highly specific labels, meaning that photographers tend to have a deeper knowledge of the subjects of their photographs and can often provide useful details, such as the model of a car or the exact species of a bird. In Figure 1[1], we see some examples of actual image pairs that can be discovered with our proposed approach and the types of highly specific tags that can be extracted from them.

The primary contribution of this work is a framework wherein tags provided by two photographers on two highly-similar images can be leveraged to find highly specific and reliable textual annotations of the visual content of the image that would otherwise be unknown to layperson annotators. We demonstrate that our approach (and similar efforts) provide more specific tags than those found using human annotators in an ESP game.

The remainder of this paper is organized as follows. In the next section, we will review techniques for providing reliable image annotations and present our new approach. In Section 3, we will present the approach in depth and then evaluate its performance on a large-scale image collection in Section 4. We review related work in Section 5 and give conclusions and future work in Section 6.

[1]All images used in this paper are licensed under Creative Commons. See Appendix A for a complete list of credits.

2. RELIABLE IMAGE ANNOTATION

In this section, we review two recent approaches that aim to increase the reliability of labeling images and propose a new hybrid approach that introduces a level of expertise and specificity to the provided labels.

2.1 Games

The ESP game [10] is a system that uses a game between two users to gather image annotations. The game relies on the assumption that agreement between two independent human annotators is sufficient for determining the reliability of a given tag or annotation. In the game, two separate players are shown the same image and asked to type in terms for objects that they see in the image. The players are awarded points if they both enter matching terms. This game interface both entices users to participate in image labeling and ensures that the collected annotations are accurate. A key criticism of this approach is that the gathered annotations will be somewhat shallow. For example, in an image containing a car, any two individual players should be able to enter "car." However, an expert might know the make and model of the car and even the year in which it was made. In the ESP game, there is little chance that this expertise will be present in both players, and this highly specific, expert knowledge will not make it into the set of trusted tags for the image.

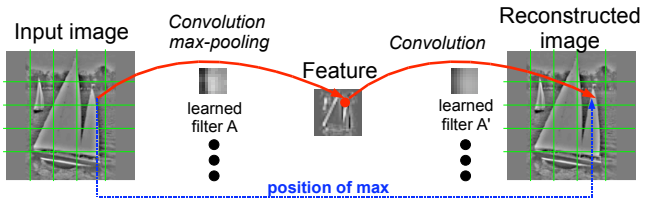

Figure 3: Block diagram of feature extraction method.

2.2 Large-Scale Image Search

One can also leverage the vast amount of images that exist in social media sites (such as Flickr™[2]), as a source of labeled image sets for learning visual models and retrieval. It has been discovered that the labels in these resources are often very noisy and not necessarily reflective of the visual content of the image. By some estimates, as few as only 50% of images with a particular tag may actually contain visual content related to that tag [5].

To address the noisiness inherent in tagged media, a number of recent efforts have undertaken an approach based on query-by-example image search [6, 7]. These methods find the set of nearest-neighbors for an image in a low-level visual feature space and then rank or weight the tags that are applied to the image based on their frequency in the set of nearest neighbors. The assumption is that, if many visually similar image share common tags with the query image, then it is likely that the given tag is highly related to the visual content of the image. If the image contains tags that are not found amongst its nearest-neighbors, then, perhaps those tags are not related to the visual content and should, therefore, be de-emphasized. The efforts to evaluate systems like these, however, have only focused on fairly common or generic objects and tags, like "beach," "boat," and "flower."

2.3 Proposed Approach

In this work, we propose a method for gathering reliably tagged images that lies between the two game- and retrieval-based approaches that we discussed above. Our primary insight is that the methods applied for retrieval-based tag selection can be leveraged to fill in the problems in expertise and specificity that are often found in the ESP game.

Specifically, photographs shared on Flickr are typically tagged by the owner of the photograph, who is usually the photographer who shot the photo. Therefore, the annotations found on Flickr are infused with a certain amount of personal context. If the photograph is a landmark or location, the photographer has most likely *been there* and can accurately name the exact spot at which he or she took the photo. Similarly, if the photograph is of a flower, bird, car, or anything else, the photographer has most likely had a first-hand encounter with that object. Since he or she finds the subject worthy of photographing, the photographer may also have a deeper knowledge of the subject. What all of this amounts to is that photographers are often more likely to be able to identify the specifics of the subjects that they photograph than random annotators. The experience of taking and sharing photographs endows the photographer with some *expertise* about the subjects of the photographs.

[2]http://flickr.com

Our proposal, then, is to examine image pairs that have high visual similarity in low-level feature space. When such pairs are found, it is similar to a case where both photographers are playing the ESP game with each other, only they are annotating slightly different versions of highly similar images. We propose that when separate authors provide identical tags for visually similar images, it is highly likely that these tags are related to the visual content of the images. We further hypothesize that these authors will be able to provide much more specific labels than other labelers, since they have a deeper personal context for the photograph at hand. In Figure 2, we show the similarities and differences between our approach and the ESP game.

3. ALGORITHM AND ANALYSIS

We have implemented an algorithm for finding reliable tags by finding identical images taken by two or more different photographers. In this section we describe the algorithm, the image-similarity feature, the calculations we performed for this paper, and the data we used to evaluate the quality of the selected tags.

3.1 Algorithm

Given a set of images, we apply a time-shifted version the ESP game. We look for image pairs that are: (1) highly similar in feature space, (2) posted by different authors, and (3) share a common tag. It is important to look for different authors because people tend to apply the same tags to many different photos (i.e. all photos from a photographer's trip to Tokyo are labeled with "tokyo." This approach is a time-shifting of the tag decision. Two photographers at different times have taken the same photo and labeled it with the same tag.

To find a set of candidate nearest-neighbor pairs for this paper, we took images in our data set and issued each as a query against the full image database. We exhaustively calculate the Euclidean distance between the query image and each image in the database to find its 20 nearest neighbors. We have implemented this approach on a large grid of computers using a Map/Reduce framework [3] and we can process thousands of query images per minute. Algorithms such as locality-sensitive hashing could also be used to speed up the search [8].

3.2 Features

The measure of image similarity can be based on many different kinds of features. In this work we use a low-dimensional feature based on a convolutional neural network.

For each image in the database, we extract a set of low-level features to encapsulate and represent the image content. The features, themselves, are learned from hundreds of thousands of web images utilizing a feed-forward network over a series of simple filters. The parameters of the network are learned to provide optimal reconstruction of images from the resulting network outputs. In the end, we are left with a 1,024-dimensional feature vector for each image, which encapsulates some aspects of the distributions of colors and textures throughout the image. Figure 3 shows a block diagram of this approach. More details are available elsewhere [4].

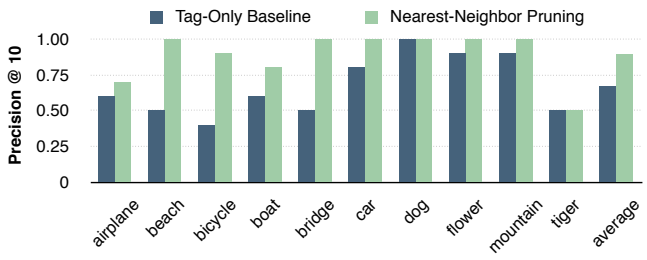

Figure 4: Precision of user-supplied tags versus reliable tags as found using the algorithm in this paper

We note that our proposed approach is not reliant on these specific features. Indeed, any combination of features that yield a reasonable measure of image similarity should be sufficient.

3.3 Experimental Data and Features

We test our proposed method on a subset of images extracted from Flickr. The set contains every publicly available photograph posted by a random sampling of 104,670 users. This totals to over 19.6 million images. In addition to the images themselves, we also gather other pieces of information around the image, such as the list of tags that have been associated with the image and the identity its owner. We normalize the tags to remove capitalization and spacing such that "Golden Gate Bridge" and "goldengatebridge" are equivalent.

For this paper we used 1 million images as queries against the full database. In our data set, roughly 7% of the query images have a neighbor that meets all of the criteria mentioned in Section 3.1. Many of these images have several such neighbors, so in total, we find more than 160,000 visually similar image pairs with matching annotations from different authors.

4. EVALUATION

We evaluated our approach to find reliable tags using three different tests. We first measured the precision of the original (photographer-specified) tags versus the precision of our reliable tags. Then we describe the specificity of the tags and categorize the types of images for which we find specific tags. Finally, we compare the specificity of the tags found algorithmically by our proposed method against those provided by random annotators in and ESP-like scenario.

4.1 Precision of Discovered Terms

A first question to ask with a set of visually similar images with shared tags is: are these tags, indeed, accurate? Other works in this area [6, 7] have conducted evaluations on the precision of re-ordered or re-weighted tags by applying them in image search scenarios. In a baseline approach, we conduct a search for a term and gather a random set of images tagged with that term. To evaluate the relative improvement in tagging accuracy provided by our proposed approach, we conduct an additional search where we constrain the set of returned images to be only those that are both tagged with the search term and also have a nearest-neighbor that is also tagged with the search term, effectively limiting the search set to only those images for which we are most certain of the accuracy of the tag.

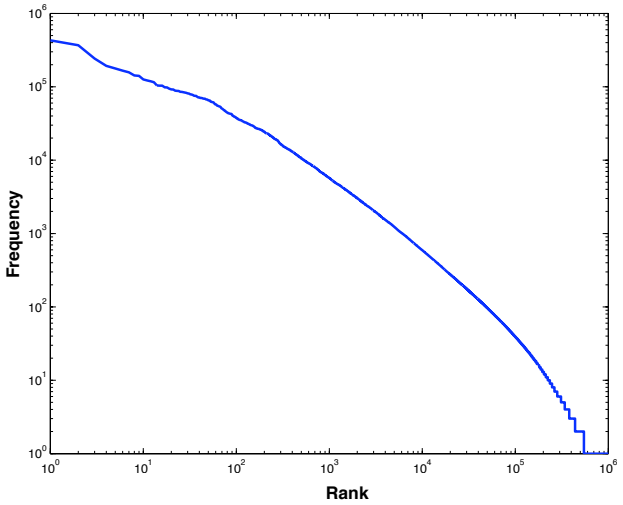

Figure 5: Tag frequency versus rank.

We conduct this evaluation over our collection of images using the 10 query terms evaluated by Li et al. [6]. The results of the evaluation are shown in Figure 4. For each query term, we calculate the precision of the top-ten returned results (the percentage of the top-ten results that are actually relevant). We see consistent improvement in search relevance across all of the search terms and a 37% average relative increase in precision. Even the false-positive images returned by our system are understandable. Many images tagged with "airplane" or "boat" were not photographs *of* airplanes or boats, but in fact photographs of the view taken *from* airplanes or boats. Similarly, many of the images tagged with "tiger" were actually of sports teams named "Tigers."

4.2 Specificity of Annotations

Having seen that the tags shared between visually similar images are sufficiently accurate, we move on to evaluate the level of specificity in the tags that we have discovered. Specificity is difficult to measure. In this study, we simply use the frequency of terms in the overall database as a proxy for estimating their specificity. The intuition is that less-frequent terms contain more information than more-frequent ones: that "San Francisco" is more specific than "California" and that "Golden Gate Bridge" is even more specific. Rareness, of course, is an imperfect measure of specificity. We might also consider employing structured knowledgebases, such as WordNet [2], to directly encode specific / general relationships between terms; however, in systems like Flickr, where users employ a large vocabulary, which is often divergent with standard English, frequency might still be the best measure available.

Our collection of images contains roughly 2.6 tags per image and more than 1 million unique tags. Figure 5 shows the frequency of each tag compared to its rank and shows something similar to a power law distribution, which is typically expected.

To begin our investigation of the quality of specific tags discovered by our method, we sort the pairs of images by the frequency of the least-frequent tag that the two share

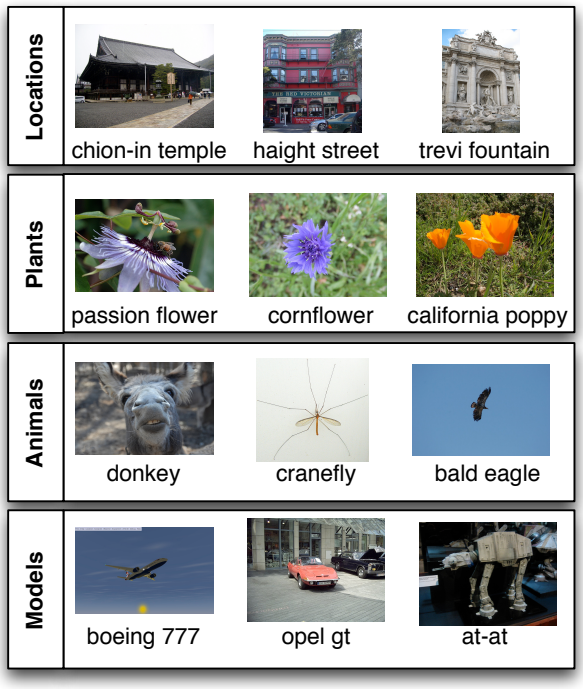

Figure 6: Examples of categories of discoverable infrequent tags. Each shown image is found automatically by our system.

and inspect the image pairs discovered that share the least-frequent tags. Upon inspecting this set, we find that these specific tags fall into four primary categories:

- **Locations and Landmarks** account for more than 50% of the discovered pairs with infrequent tags. These include specific city names as well as specific sites, such as "san francisco" and "coit tower."
- **Plant-life** photographs account for approximately 25% of the discovered pairs. These are often photographs of flowers with correct names attached, such as "dahlia" or "orchid."
- **Animal-life** accounts for about 10% of the discovered pairs and includes specific species of insects and animals, such as "crane fly" and "common kingfisher."
- **Makes and Models** include tags related to specific products, such as "audi a3" and "zune." These are found in about 15% our examples.

Examples of each of these specific tag categories can be found in Figure 6.

4.3 Human vs. Algorithmic Specificity

To further evaluate the level of specificity of the tags that we have discovered, we compared the specificity of human and machine-generated tags. We conducted a simulation of the ESP game using some of these specifically-tagged image pairs. We select 100 image pairs, all with rare shared tags, where rare is defined as tags that occur in less than 0.005% of the images in our database. These contain locations, plants, animals, and makes/models in roughly the same proportions

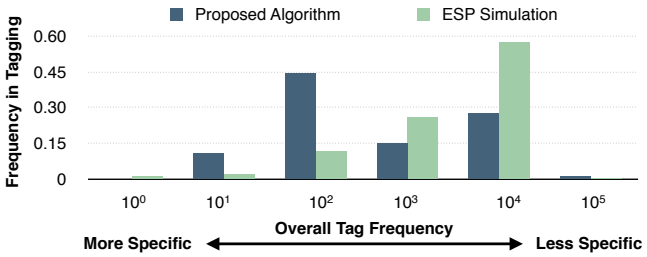

Figure 7: Distribution of specificity of tags provided by human annotators in an ESP game simulation versus the tags provided by photographers and discovered algorithmically by our approach. Less specific tags, those with high-frequency on the right of the graph, are used more often by human annotators (in an ESP game), compared to the specific tags on the left.

as discussed above. We remove some noisy pairs that were found due to non-English tags, or tags related to web applications used to make certain types of images or collages. We show one image from each pair to two subjects, both of whom are native English speakers, residents of San Francisco, and unfamiliar with the hypotheses and methods that we have employed in this work. The subjects are both shown the same image from the pair and are asked to provide a set of descriptive terms related to the image content, so they are effectively asynchronously playing the ESP game.

In Figure 7, we show the distributions of the specificity of the tags found algorithmically for the image pairs (the tags provided by the photographers) and the tags provided by the image annotators in our study. We see a stronger tendency in the human annotators to provide generic tags (those that are generally more frequently used), while the algorithmically discovered tags were more specific, overall.

Further analysis of the study results are shown in Table 1. For each image, each annotator provided, on average, around 3 tags. First, we counted how many times both annotators had at least one tag in common. This was a fairly common occurrence, with matches found on 78% of all the images, which reinforces the utility of the ESP game in general: it works and provides reasonable annotations. We then count the number of times that the human annotators had at least one *specific* tag in common. (Here, we counted any tags that referred to specific locations, species of plant or animals, and brands or models.) Specific matches occurred less frequently.

For animals and models, the human annotators had a reasonable degree of success, though many of these were due to identifying the brand of a car where the logo was clearly visible in the photograph or identifying somewhat common animals, such as bees and donkeys. In no cases, however, were the annotators able to identify the exact model of a car. More-obscure animals, such as the common kingfisher and the crane fly, also presented some difficulty.

The performance of the human annotators on the location images was more mixed. They were able to identify a number of well-known locations and landmarks, like the Liberty Bell and the Statue of Liberty. However, the annotators did not have enough knowledge for some more obscure locations, like the Harbour Bridge in Sydney or Monument Valley.

Plants were very difficult for the human annotators. These

Category	# Images	% Agreement	% Specific
Locations	54	67%	20%
Plants	24	92%	4%
Animals	10	80%	40%
Models	13	92%	46%
Total	100	78%	22%

Table 1: The quality of human-generated tags in an ESP-like game on the 100 evaluated image pairs for which this algorithm found highly specific reliable tags. For each category, we show the total number of images annotated (# Images), the percentage for which the two annotators provided a matching tag (% Agreement), and the percentage of pairs for which the annotators were able to provide highly specific tags (% Specific).

tended to include a number of species of flowers. The annotators were able to correctly identify an orchid, but were unable to identify marigolds, corn flowers, California poppies, or any other type of plant.

Interestingly, when the human annotators tried to provide some specific information, the tags that were entered were often erroneous. For example, one annotator mistakenly identified a building by architect Frank Gehry to be by Antoni Gaudi and entered the incorrect model of a particular Audi car. The other annotator misidentified a building in Barcelona as being in Germany. And this actually, again, emphasizes the power of the ESP game: such erroneous input would be ignored, since it is not confirmed by the other player. However, these results also suggest that if the annotators are also the photographers and they have some degree of expertise on the subject of their photographs (as is the case on Flickr), then it is possible to mine some highly specific annotations from their shared photo collections. This approach may exceed the capabilities of the standard ESP game with respect to gathering annotations that are reliable and specific.

Our proposed method, on the other hand, leverages the personal context of the photographers who have actually taken the photographs and is able to discover tags for images that are often highly reliable and very specific.

5. RELATED WORK

Aside from the works that we discussed in Section 2, there are some other related works that have aimed to leverage query-by-example image search over large collections of images to provide image annotations. Crandall et al. [1] and Zheng et al. [11] have both proposed systems that work over large collections of geo-referenced photographs. Given a query image, the systems can find visually similar images from the collection, identify the location at which the photograph was taken, and provide tags for the subject of the photo. Torralba et al. [9] constructed a collection of 80 million images from the web image search results for every term in WordNet. They then proposed to label unseen images by finding their nearest neighbors in the search set and propagating the labels from those images to the query image. These works are differentiated from our work in that they seek to predict labels for untagged images, whereas we have focused on mining knowledge from pairs of images that are both labeled by human annotators.

6. CONCLUSIONS AND FUTURE WORK

Tags are inherently a noisy, user-generated signal describing the content of an image. We wish to find words that multiple people can agree apply to an image in hopes the process will yield tags that are *reliable* and *specific*.

In this work, we demonstrated a novel way to generate highly reliable tags that describe an image. Unlike the ESP game which asks two random users to find a tag they agree describes a photo, we look for highly similar images that are described with the same term by two different photographers. The most important benefit of this approach is that we often find highly specific tags, which give a greater level of detail about the subjects of the images. This specificity arises out of the photographers' knowledge of the location of the object, or its exact description. In addition, recruiting new game players can be difficult and we get our tags from the words that a photographer has already supplied.

In the experiment described here, we found reliable tags for 1% of the images in our database. We tested the images for which we found highly-specific tags using our algorithm and found that only 22% of the same images were annotated in an ESP game by humans with specific tags. Furthermore, we saw a tendency for human annotators to provide more-generic annotations than the photographers, who provided highly-specific annotations. In addition, we found an average increase in the precision of the tags of 37%.

Our approach is efficient because we capitalize on the labels already provided by users. In many cases the photographer supplies the labels as part of a submission to a photo-sharing site. But we can also apply the same approach to anchor text or other captions associated with a web page.

While we applied this approach to only 19 million images or less than 1% of the publicly available images on Flickr, the success of our approach will only grow as we expand the database to cover more of the photos available on the web. We do not expect to be able to find reliable tags with every image—some images are too unique to ever find a match. But with the wealth of photos on the web, we do not need to label each image, just find good images that are tagged with any particular word. Perhaps most importantly, our approach is not limited to finding English tags.

7. REFERENCES

[1] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 761–770, New York, NY, USA, 2009. ACM.

[2] C. Fellbaum et al. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.

[3] Hadoop. http://hadoop.apache.org/core/.

[4] E. Hoerster, M. Slaney, M. Ranzato, and K. Weinberger. Unsupervised image ranking. In *Proceedings of the ACM Workshop on Large-Scale Multimedia Retrieval and Mining*, Beijing, China, October 2009.

[5] L. Kennedy, S.-F. Chang, and I. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 249–258, 2006.

[6] X. Li, C. G. M. Snoek, and M. Worring. Learning tag

relevance by neighbor voting for social image retrieval. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pages 180–187, Vancouver, Canada, October 2008.

[7] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 351–360, New York, NY, USA, 2009. ACM.

[8] M. Slaney and M. Casey. Locality-sensitive hashing for finding nearest neighbors [lecture notes]. *Signal Processing Magazine, IEEE*, 25(2):128–131, 2008.

[9] A. Torralba, R. Fergus, and W. T. Freeman. Tiny images. Technical Report MIT-CSAIL-TR-2007-024, Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 2007.

[10] L. von Ahn and L. Dabbish. Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

[11] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. Chua, and H. Neven. Tour the World: building a web-scale landmark recognition engine. In *CVPR '09: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

APPENDIX

A. PHOTO ATTRIBUTIONS

All photographs used in this paper are licensed under Creative Commons. Photographer credits are listed below.

Figure 1:
http://www.flickr.com/photos/hmorandell/130536654/
http://www.flickr.com/photos/thruhike98/248608444/
http://www.flickr.com/photos/jvverde/2298487209/
http://www.flickr.com/photos/16134053@N00/266023301/

Figure 2:
http://www.flickr.com/photos/redneck/2363234517/
http://www.flickr.com/photos/davidthibault/2449284848/

Figure 6:
Locations
http://www.flickr.com/photos/aranmanoth/350381473/
http://www.flickr.com/photos/plyn4lf/121479650/
http://www.flickr.com/photos/63269749@N00/2152878314/
Plants
http://www.flickr.com/photos/pdc/553336027/
http://www.flickr.com/photos/dickey/113608130/
http://www.flickr.com/photos/ericinsf/52601149/
Animals
http://www.flickr.com/photos/vickispix/225669748/
http://www.flickr.com/photos/96619357@N00/201296878/
http://www.flickr.com/photos/georgehoffman/168099747/
Models
http://www.flickr.com/photos/rduffy/164891469/
http://www.flickr.com/photos/volodimer/405241836/
http://www.flickr.com/photos/deep-resonance/484983237/