Aggregating Visual Evidence from Social Media Photos to Monitor the Natural World

ABSTRACT

Social photo-sharing websites collect a huge amount of latent visual information about the world, including information about the environment and ecology. In this work, we propose to reconstruct satellite maps of environmental status across North America through millions of publicly available geo-temporal tagged images. We apply modern deep learning-based recognition techniques to identify phenomena in images, and then aggregate evidence from multiple users to estimate whether or not the phenomena were occurring in a given time and place. We then evaluate the accuracy of these estimates by comparing to actual satellite maps as ground truth. As test cases, we consider two important ecological phenomena for which high quality ground truth is available: snowfall coverage and vegetation (greenery) coverage. We find that while the automatic recognition techniques are noisy on any single particular image, we can accurately estimate the phenomena's presence when enough users have uploaded enough photos at a particular time and place. This evidence from photo-sharing websites could create new sources of data for ecologists, perhaps helping to overcome the limitations of traditional data collection techniques like manual observation (which is labor intensive) or satellites (which are not able to observe through clouds).

Keywords

ACM proceedings; LaTeX; text tagging

1. INTRODUCTION

Monitoring the meteorology and vegetation phenomenon is the cornerstone and key challenge of ecology and biology research. Expensive satellite images give large scale data but are limited by cloud cover, atmospheric conditions, struggle with fine-grained localization such as flower species distribution and are not applicable on observing human interaction with nature; while citizen science provides high quality data but is also costly and is very difficult to practice over large scale areas. The enormous popularity of photo-sharing website collects images in large spatial scale, from under clouds and in close focus (compare to aerial surveillance), moreover, they are freely accessible to the public. The more than 300 million images uploaded to social media every day [1] potentially contain not only human activities, but also outdoor ecology and biology information intentionally and incidentally as shown in Figure 1.

Figure 1: Flickr image examples capture snow and greenery evidence on purpose and as background.

The idea of reproducing satellite maps has become more and more interesting to scientists applying textual mining on [2, 21, 20, 19] and recently to computer vision researches directly deriving various information directly from visual contents such as temperature [6], dynamic status of cloud [11], snow coverage on mountain peak [4, 5]. In this paper, we test the feasibility of leveraging these noisy and biased images as a new source to observe nature. We study 2 particular phenomena, snowfall and vegetation coverage as they are fundamental topics in ecology and biology study, have relatively distinct appearance to recognize, have a good chance to appear in social media, and also have satellite maps available to serve as ground truth. Our approach is illustrated in Figure 2. First, we collect a large hand-labeled data set of the existence or absence of ecology phenomena. Then, we train a classifier for each phenomenon by combining its most discriminative visual features and by using deep learning features. Finally, we collect 12 million images from entire North America over 4 years, make prediction on geo and temporal scale by aggregating visual evidence.

Inspried by an earlier work [21] analyzing ecology phenomenon from image tags only. We apply a new approach by understanding visual content of images, and run experiments on the exact same data set to study how vision techniques could help in social media data mining compared to using textual data alone. Also, to our best knowledge, among all the research works performing social sensing with image data, this is the first one providing continental scale quantitative performance evaluation.

2. RELATED WORK

In last few years, crowd-sourcing data from social media

DOI: 10.1145/1235

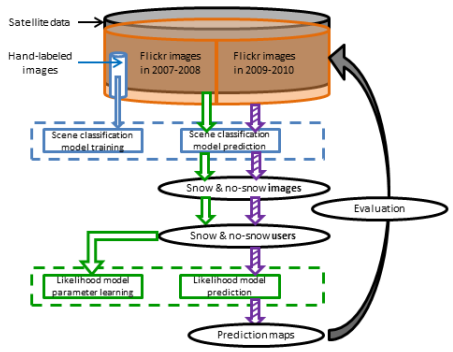


Figure 2: Overview of our approach. We train image classifiers(in blue) on large scale images. And by applying it to training images in 2007-2008, we train a likelihood model(in green) and finally make prediction by aggregating these visual evidence.

as a large scale and free to public data source has **FiXme Note: received lots of attention from; or (become more and more popular to)** researchers working on using textual contents to predict elections [20], using geo-tags to quantify tourism in nature area using geo-tag profile of social media users [19], **FiXme Note: talk a lot about motivation of scientific report paper since it's in nature area** to draw coastline [15], using geo and temporal tags to analyze people's event-based activity when large group of people gathering together during a function of time such as football match, and using both geo and textual tags to extract land use information from Panoramio [17, 13], in [21], Zhang *et al.* estimates snowfall and vegetation coverage based on geo, temporal and textual tags of Flickr images.

Since public-sharing photos provides such a huge potential in social and environmental study, it's natural to see a lot of works start analyzing image contents. Webcam providing dense temporal images is a good source to monitor the nature. A series of works explore sequences of webcam images describing outdoor scene with 40 transient attributes [8], estimating dynamic cloud maps [12, 11], exploring interactions between visual elements and the temperature **FiXme Note: or just as the title: exploring correlations between appearance and temperature** [6], and monitoring the dynamic snow phenomena at mountain areas [4, 5]. To evaluate the study of temperature, cloud, and snowfall amount, researchers can easily compare their results with satellite maps. Some works also use crowd-sourcing data from other sources, for example, Google street view provides selectively dense geo distributed images to help navigating the environment [7] and understanding urban scene and predicting urban perception [16], and Li *et al.* use the co-occurrence statistics of celebrities appears on news images to auto tag photographs of celebrity community [9] Unfortunately, the evaluation in these works are either not in continental scale or just via quality visualization. Performance of social activity studies, on the other hand, are even harder to evaluate. **FiXme Note: say more about our evaluation? or move this to another place?**

Flickr and Panoramio as very popular photo-sharing websites "involuntarily" support researchers identifying salient city attributes and analyzing the visual similarity among different cities in order to apply computer vision to urban planning [22] Photo-sharing websites collecting visual contents directly from people's activity and their surrounding areas which is so important, hard to collect otherwise but also very noisy. **FiXme Note: write something about so there are very few work appears and so we are working on this?**

The fact that webcam can only be placed far away from people makes it almost impossible to monitor people's activity, even not the surrounding area close to residential or **FiXme Note: crowd? I mean groups of people like downtown, not ski activity but like people going to work and back everyday also a good topic to use temporal dense images but Webcam is not good at this.** Social media, on the other hand, provides a larger freedom on location distribution. In fact, as a complementary, almost all the photos shared online are from locations people usually go to. **FiXme Note: how helpful is this to study more areas close to urban planning, market sharing, everyday living, anything related to people**

Our work take the advantage of studying ecology phenomena with **FiXme Note: easy to get, more reliable** satellite maps as ground truth and use social media data to **FiXme Note: monitor? insight?** these information from **FiXme Note: locations more related to people**. We provide continental scale quantitative evaluation and introduce our method to tackle the problem of noisy and biased data, in order to support extended studies in other areas. **FiXme Note: just want to say more areas in natural or not only natural but also social**

3. EXPERIMENTS AND RESULTS

In this paper, we propose an effective system to estimate the presence or absence of a given ecology phenomenon (like green leaf plants) from visual contents of public-shared photographs. First, we employ the state-of-the-art computer vision technique to detect the target phenomenon. Then according to the visual evidence and the corresponding timestamp and geo-tag of each image, we adopt the likelihood model in [21] aggregating the large scale, imperfect information to reconstruct the satellite maps.

In Scene classification, we also employ the current state-of-the-art algorithm, the Convolutional Neural Network (CNN) pre-trained on ImageNet dataset. The key idea behind this approach is that instead of selecting and combining hand designed features with limited parameters to fit each recognition task, it learns hierarchical image feature directly for target objects. We fine-tune it with our hand-labeled dataset for each experiment case and further improve the performance illustrated below.

We study two important ecology phenomena, for each time period and location, whether there is snowfall and are there many plants with green leaves? We discuss the experimental setup and the results and evaluation below.

3.1 Snow Case

3.1.1 Scene Classification

Beyond combining visual features, we build CNN visual model for our snow scene classification problem by fine-tuning Imagenet pre-trained model with the 8000 training images labeled for snow scene [18]. The best performance in [18] using visual features with SVM is 80.50% accuracy.

On the same testing set of 2000 images, our CNN model achieves 88.06% accuracy as a significant improvement. Details of comparison between performance of CNN model and other visual features are presented in Table 1. Therefore, we use this as our recognition model for final predictions.

Figure 3 compares classification performance of CNN with individual visual features and their combination in terms of an ROC curve, as well as a precision-recall curve in which the task is to retrieve photos containing snow. The precision-recall curve of CNN shows that at about 50% recall, precision is very close to 100%, while even at 80% recall, precision is still above 90%. This is a nice feature because in many applications, it may not be necessary to correctly classify all images, but instead it is important to find some images that most likely contain snowy scene.

We now turn to present experimental results for estimating geo-temporal distributions of snowfall.

3.1.2 Snow Prediction on Cities

To compare with existing results in [18] using tag based method, we first test how well our image-content based method can predict snowfall on daily base at a local scale, and in particular for the same four U.S. metropolitan areas, New York City, Boston, Chicago and Philadelphia. Table 2 shows some basic statistics for these 4 cities, and results of these classifiers. Best performance obtained when we combine the confidence scores of tags and visual model based on CNN. For each of the method, Chicago gives the highest accuracy while Philadelphia gets the lowest accuracy. It's reasonable considering that Chicago has the most active Flickr users per day (94.9) while Philadelphia has the least (43.7).

There is not a clear evidence that visual evidence is more informative in estimating snowfall presence. In contrast, it gets lower accuracy in all these 4 cities. But combining tag and visual confidence achieves considerable improvement in performance. We apply this combined model in the following experiments.

3.1.3 Continental-scale Snow Prediction

Adding visual evidence, we reconstruct Satellite maps and evaluate our model in continental scale. We follow the same metric in [21] to produce estimation at resolution of 1×1 degree (roughly $100 \times 100 km^2$) square and get one map for each day in 2009 and 2010.

Figure 4 shows the precision and recall curve of snow and non-snow prediction in continental-scale. Here we limit our predictions for the bins which both have photos taken at that time and location and have satellite ground truth available. We computed our confidence scores based on tags and image-classification, then we trained simple decision tree to learn the correct thresholds to make final prediction. We achieve almost 0.5% over the baseline (cutting the error rate by more than 20%), the baseline in our case is the majority class which predicts non-snow all the time.

3.2 Vegetation Case

Vegetation coverage is a more stable phenomenon, so we study this case in every 16 days period instead of daily base. Unlike snow scene that tag feature is very discriminative, according to [21], textual tag of greenery is highly misleading. Terms like "tree" or even "leaf" are not necessarily indicating green color, but terms like "green" or "yellow" could be used to describe a wild set of objects outside of vegetation.

Table 1: Performance of different features for snow detection using SVMs for classification and compared with CNN model.

Feature	Kernel	Accuracy
Random Baseline	—	50.0%
Gist	RBF	73.7%
Color	χ^2	74.1%
Tiny	RBF	74.3%
Spatial Color Moments	RBF	76.2%
Spatial pyramid LBP	RBF	**77.0%**
All features	linear	**80.5%**
CNN	-	**88.1%**

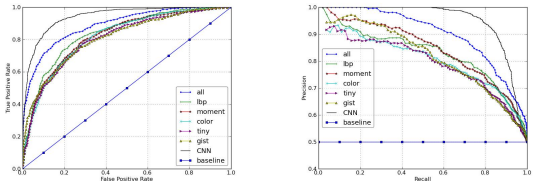


Figure 3: Snow classification results for different features and combination, in terms of *(left):* ROC curves for the task of classifying snow vs. non-snow images; and *(right):* Precision-Recall curves for the task of retrieving snow images.

Table 2: Selected basic statistics during 2007 to 2010 for the 4 cities and results of the likelihood model using tags and vision evidence.

City	naive baseline	active user/day	snow days	tag con↓dence	vision conf	tag conf & vision conf
NYC	85.00%	65.6	185	90.42%	90.29%	92.34%
Chicago	72.80%	94.9	418	94.12%	93.16%	95.08%
Boston	75.60%	59.7	373	89.18%	85.21%	91.23%
Philly	80.50%	43.7	280	89.19%	85.09%	89.19%

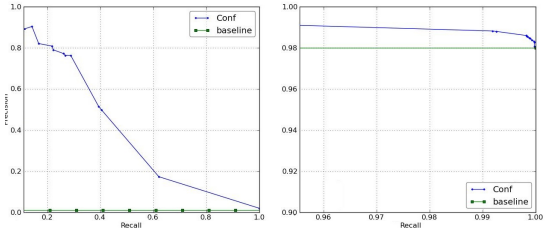


Figure 4: Precision and recall curve of snow prediction (left) and nonsnow (right) in continental scale.

Moreover, while vegetation is such a common natural scene appears in most of the outdoor images, it is very unlikely people would include it in their textual tags to describe the image. In this case, from a huge number of images contain greenery shared online, only a very small portion of them actually provides related textual tags. On the other hand, we believe visual feature can be very descriptive and exclusively describing vegetation with green leaves.

3.2.1 Dataset

We build a data set with over 10000 Flickr images taken before 2009, and are composed by images with "forest" and "summer" like tags and also random images without any tag preference. These images are hand labeled with categories *"Outdoor Greenery","Outdoor non-Greenery","Indoor","Other-modified"*,and *"Not available"*.

Finally, we build a positive set with images in category *"Outdoor Greenery"* and a negative set with images in categories *"Outdoor non-Greenery"* and *"Indoor"*. To learn a image classification model, we build a training set with 4000 images and a testing set with 1900 images. In training and testing set, there are equal number of positive and negative samples. To show the diversity of our Flickr image dataset, in figure 5 we present a random sample of images in our vegetation dataset labeled as positive and negative.

In continental-scale prediction, we only look at images on Flickr.com in year 2007 to 2010, with no tag limitation. We filter out the photos with too unreasonable timestamps such as the taken time and uploading time is exactly the same. We only use images with high precision geotag according to the image metadata. And we still process the satellite ground truth the same way as in [21].

3.2.2 Scene Classification

We hand-labeled a separate dataset for greenery scene. Compare to snow scene recognition, vegetation has the signature green color that the biologists are very interested in during exactly which time period they are green. Thus it's very important to find out when it turns from yellow to green and when does it turns back to yellow. The leaves of plants also have distinctive visual texture. Thus, visual features capture both color and texture information are our best choice. So we employ color SIFT feature to analyze the local gradient distribution. And we also extract color GIST feature to describe texture feature and global context.

Color SIFT histogram. We extract dense SIFT feature [10] on each of the RGB color plane, and concatenate them to build color SIFT feature. The dense SIFT feature is extracted from every 2 pixels by 2 pixels bin, with a step size of 5 pixels. In this way, we achieve representative key points and reasonable computation complex.

From training data set, We build 2000 dimensional centers of color SIFT feature using K-means clustering. With these centers, a 2000 dimensional histogram is built from all the key points of each image.

Color GIST. Similarly, we also extract GIST features on RGB color channel respectively. **FiXme Note: From now on to the end of this paragraph, it's the same as our workshop paper. Could you help me to give a shorter global description?** GIST feature capture coarse texture and scene layout by applying a Gabor filter bank followed by down-sampling [14]. Our variant produces a 1536-dimensional vector and operates on color planes. Scaling images to have square aspect ratios before computing GIST improved classification results significantly [3].

By concatenating color SIFT histogram and color GIST feature, a model is trained and tested with SVM using RBF kernel. It achieves accuracy of 85.90% though CNN still gives a higher performance of 88.00%. We present detail performance in Table 3.

3.2.3 Vegetation Coverage over Time and Space

We consider North America area as in snow experiments. From images in 2007 and 2008, we learn the prior probability of a place being covered by vegetation at any given 16-days period is 75.16% For any user taken photos from a place covered by greenery at a given time, the probability of the photos contain greenery scene is 27.18%. On the other hand, there is only a small chance 3.03% to see a user uploading images with greenery in a place not covered by enough vegetation according to satellite observation. Since residence enjoy meadow around their house and it is unlikely not to see any plants wherever people would go, there is a considerable chance to see green scene in photos, but when there are enough users uploading images, we will still be able to distinguish the actual green and non-green area.

While the satellite has ground truth for 87594 bins in North America, our method predicts our method predicts 61602 bins (70.3% in quantity). Moreover, about 20% of satellite ground truth locate in north Canada where the ecology system is stable and very little human-environment interaction happens. Moreover, our data is from users in social media. So our prediction focus on more populated locations or places people like to visit such as natural scenery.

For evaluation purpose, we only evaluate the time and location both Flickr and Satellite have data in North America. The overall accuracy of our method is 93.2% comparing to the 86.6% majority baseline. The precision of green bins is 98.8% and the precision of non-green bins is 68.2%. Recall of green bins is 93.3% and recall of non-green bins is 92.5%. In [21], the performance of predicting vegetation coverage of tag feature is not going to improve the result of using visual evidence. Figure 6 shows the precision and recall curve of vegetation prediction in continental-scale.

Generally, almost all the false negative error is due to the sparseness of data. While not enough images are collected at certain location during some time, there is either no green image found or green images are too few compare to the quantity of non-green images. On the other hand, false positive error is rare (less than 1%) and complex. We found most images in the false positive bins are actually green vegetation images. **FiXme Note: here we need some more explanation** In figure 7, we show some examples of images in false positive bins.

3.2.4 Performance at single place over time

For each single location, we can find out when exactly did the leaves turn yellow as well as when did the leaves turn back to green. Figure 8 shows vegetation coverage of 6 places over 2009 and 2010. Prediction results on top usually have more data available than ground truth on the bottom. And we can see the ground truth is likely missing at the time when the leaves change color.

3.2.5 Single time over places

Sample maps are presented in figure 9. These maps are visualization of the performance in North America. We use public sharing Flickr images suffering sparsity in locations, but are more likely taken from more populated or more popular locations. The satellite ground truth, instead, is limited to unpredictable cloud coverage and other sensor precision issue.

Table 3: Results of our visual models for vegetation.

Visual feature	Accuracy
Random Baseline	50.00%
Color SIFT	78.10%
Color GIST	82.58%
SIFT and GIST	85.90%
CNN%	88.00%

(a) Random positive images in vegetation dataset

(b) Random negative images in vegetation dataset

Figure 5: Random images from our hand-labeled dataset. Public sharing images are various in quality, contents, illumination and view angle. Negative images like winter trees without leaves, or indoor images capturing a photo of forest are more confusing.

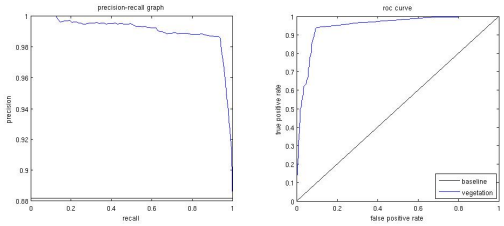


Figure 6: Precision and recall curve of vegetation prediction in continental scale.

Figure 7: In vegetation detection over North America in 2007, among all false positive bins, there are 47 images that are predicted as greenery. And they are the reason these bins are predicted as green. FiXme Note: conclusion of very green images in non-green bin

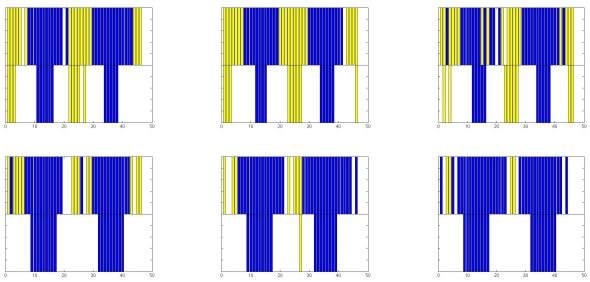
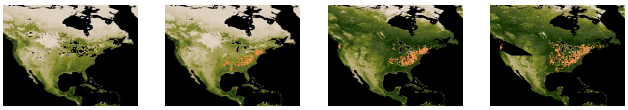
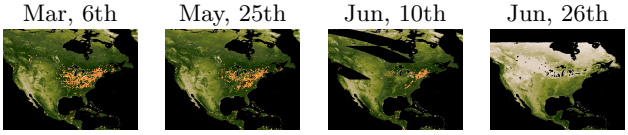


Figure 8: Yellow bars show non-greenery at that time. Blue bars represent greenery. Prediction results on top shows 6 random places comparing to satellite ground truth. The ground truth on the bottom tends to disappear when leaves are turning yellow or green.



Mar, 6th | May, 25th | Jun, 10th | Jun, 26th



Aug, 13th | Aug, 29th | Sep, 14th | Dec 19th

Figure 9: We use prediction results to recreate vegetation coverage maps for each 16-days period. There are 8 maps picked in 2010. The dates under each map are the starting date of each 16-days period. Orange bins represent true positive; blue bins are false positives; yellow bins are true negatives and black bins are false negatives.

4. REFERENCES

[1] top-15-valuable-facebook-statistics. http://www7.ncdc.noaa.gov/IPS/cd/cd.html.

[2] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[3] M. Douze, H. Jegou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *ICIVR*, 2009.

[4] R. Fedorov, P. Fraternali, C. Pasini, and M. Tagliasacchi. Snowwatch: Snow monitoring through acquisition and analysis of user-generated content. *arXiv preprint arXiv:1507.08958*, 2015.

[5] R. Fedorov, P. Fraternali, and M. Tagliasacchi. Snow phenomena modeling through online public media. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 2174–2176. IEEE, 2014.

[6] D. Glasner, P. Fua, T. Zickler, and L. Zelnik-Manor. Hot or not: Exploring correlations between appearance and temperature. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3997–4005, 2015.

[7] A. Khosla, B. An, J. J. Lim, and A. Torralba. Looking beyond the visible scene. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3710–3717. IEEE, 2014.

[8] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (TOG)*, 33(4):149, 2014.

[9] L.-J. Li, D. A. Shamma, X. Kong, S. Jafarpour, R. Van Zwol, and X. Wang. Celebritynet: A social network constructed from large-scale online celebrity images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 12(1):3, 2015.

[10] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[11] C. Murdock, N. Jacobs, and R. Pless. Webcam2Satellite: Estimating cloud maps from webcam imagery. In *WACV*, 2013.

[12] C. Murdock, N. Jacobs, and R. Pless. Building dynamic cloud maps from the ground up. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 684–692, 2015.

[13] H. Oba, M. Hirota, R. Chbeir, H. Ishikawa, and S. Yokoyama. Towards better land cover classification using geo-tagged photographs. In *Multimedia (ISM), 2014 IEEE International Symposium on*, pages 320–327. IEEE, 2014.

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[15] M. Omori, M. Hirota, H. Ishikawa, and S. Yokoyama. Can geo-tags on flickr draw coastlines? In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 425–428. ACM, 2014.

[16] L. Porzi, S. Rota Bulò, B. Lepri, and E. Ricci. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 139–148. ACM, 2015.

[17] M. Šećerov. *Analysis of panoramio photo tags in order to extract land use information*. PhD thesis, 2015.

[18] J. Wang, M. Korayem, and D. Crandall. Observing the natural world with flickr. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 452–459, 2013.

[19] S. A. Wood, A. D. Guerry, J. M. Silver, and M. Lacayo. Using social media to quantify nature-based tourism and recreation. *Scientific reports*, 3, 2013.

[20] Q. You, L. Cao, Y. Cong, X. Zhang, and J. Luo. A multifaceted approach to social multimedia-based prediction of elections. *Multimedia, IEEE Transactions on*, 17(12):2271–2280, 2015.

[21] H. Zhang, M. Korayem, D. Crandall, and G. LeBuhn. Mining Photo-sharing Websites to Study Ecological Phenomena. In *WWW*, 2012.

[22] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *Computer Vision–ECCV 2014*, pages 519–534. Springer, 2014.