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 challenge of ecology and biology research. Expensive satellite images give large scale data but struggle with cloud cover, atmospheric conditions and fine-grained localization such as flower species distribution, 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 [**?**] potentially contain not only human activities, but also outdoor ecology and biology information intentionally and incidentally as shown in Figure 2.

The idea of reproducing satellite maps has become more and more interesting to scientists applying textual mining on **FiXme Note: citestock, ecology, election, tourists**, and recently to computer vision researches directly deriving **FiXme Note: citetemperature, cloud, mountain peak** information from visual content. In this paper, we test the feasibility of leveraging these noisy and biased images as a new approach to observe nature. We study 2 particular phenomena, snowfall and vegetation coverage as they are fundamental topic 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 1. 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 2 years, make prediction on geo and temporal scale by aggregating this visual evidence.

This paper is built on our earlier work **FiXme Note: citewww** analyzing ecology phenomenon from image tags only. We apply a new approach 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 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 [**?**], using geo-tags to quantify tourism in nature area using geo-tag profile of social media users [**?**], **FiXme Note: talk a lot about motivation of scientific report paper since it's in nature area** to draw coastline [**?**], using geo and temporal tags to analyze people's event-based activity when large group of people

DOI: 10.1145/1235

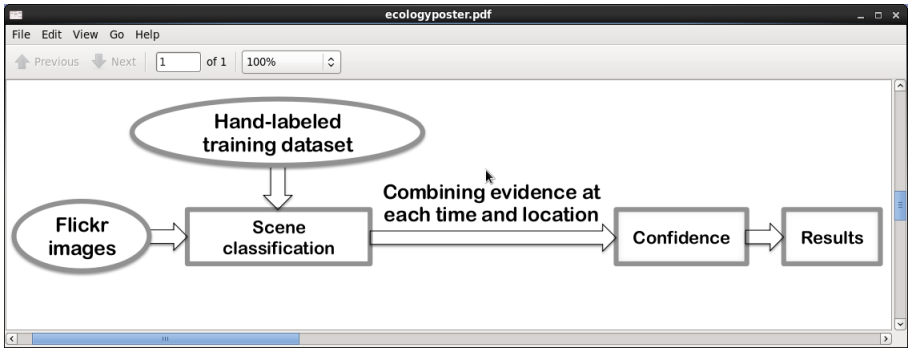

Figure 1: Overview of our approach to apply image classifiers on large scale images and make prediction by aggregating these visual evidence. FiXme Note: first classifier, then prediction

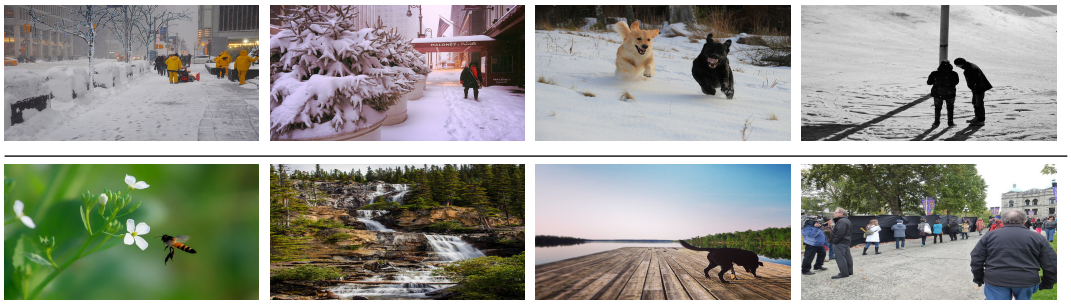

Figure 2: Flickr image examples capture snow and greenery evidence on purpose and as background.

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 [?, ?], in **FiXme Note: citewww**, Zhang *et al.* estimates **FiXme Fatal: missing letter in compiling** snowfall and vegetation coverage based on geo, temporal and textual tags of Flickr images. **FiXme Note: accuracy of geo-temporal problem, and now it's getting better.**

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 [?], estimating dynamic cloud maps [?, ?], exploring interactions between visual elements and the temperature **FiXme Note: or just as the title: exploring correlations between appearance and temperature** [?], and monitoring the dynamic snow phenomena at mountain areas [?, ?]. 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 [?] and understanding urban scene and predicting urban perception [?], and Li *et al.* use the co-occurrence statistics of celebrities appears on news images to auto tag photographs of celebrity community [?] **FiXme Note: give a term like social identity?**. 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 [?] 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?**

FiXme Note: not sure about keeping this paragraph or not. it seems duplicate of the 2 paragraphs above – 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 phe-

nomena 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. METHOD

3.1 Overview

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 use a likelihood model aggregating the large scale, imperfect information to reconstruct the satellite maps.

We collect 12 million photos shared on Flickr during 2007 and 2010 with geo-tags and timestamp available, in North America continent similar to **FiXme Note: citewww**, and split them into a training set D_1 for prior knowledge learning with photos taken in 2007 and 2008, and a testing set D_2 with photos taken in 2009 and 2010. In the two steps of training stage, for snowfall and greenery respectively, we first prepare a hand-labeled dataset similar to **FiXme Note: citeworkshop** sampled from photos taken before 2009, no matter it has geo or temporal information or not. Half of the photos must have tags indicating snow or greenery presenting in the image as positive samples, and the other half are random negative samples. We learn a scene classification model from this hand-labeled dataset for each target phenomenon. By applying this model to the continental scale dataset D_1, in the second step of training, we learn the parameters for the likelihood model introducing later in **FiXme Note: section 3.3** to make estimation for each pixel on our result map. We then test the entire system with the separate continental scale dataset D_2 and compare the results with satellite maps described in **FiXme Note: citewww** as ground truth for evaluation.

3.2 Scene Classification

3.2.1 Combining Visual Features

To train a scene classification model for snow scene and greenery scene respectively, we start with combining the most discriminative image features. Similar to snow scene recognition in **FiXme Note: citeworkshop**, 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 **FiXme Note: citeSIFT** 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 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 **FiXme Note: citeoliva2001modeling**. 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 **FiXme Note: citedouze2009evaluation**.

By concatenating color SIFT histogram and color GIST feature, a model is trained and tested with SVM using RBF kernel.

3.2.2 Deep learning

Recently the Conventional Neural Network (CNN) **FiXme Note: citekrizhevsky2012imagenet** has gained a lot of attention in the vision community, as it outperformed all other techniques in the ImageNet challenge (the most famous object category detection competition) **FiXme Note: citeilsvrcarxiv14**.

CNN is currently the state-of-the-art algorithm of image classification on standard datasets. CNN enjoys additional features that distinguish it from the standard neural networks: shared weights and sparse connectivity. A layer in CNN may consist of three different stages: convolution, nonlinear activation, and pooling. In the convolution stage, a set of convolution filters is applied in parallel. The output of a convolution filter is then passed to non-linear activation functions (e.g., rectified linear activation function, sigmoid activation function). The final stage is pooling, where the net output is manipulated based on its neighbors (e.g., max pooling, L_2 norm, and weighted average). Pooling makes the network invariant to the translation of the input

The key idea behind this approach is that instead of first designing low-level features by hand and then running a machine learning algorithm, a single unified algorithm should learn both the low-level features and the high-classifier simultaneously.

We apply CNN to detect snow and vegetation on image level. We followed **FiXme Note: citeOquab14** and started with a model pre-trained on the huge ImageNet dataset then we train our models using hand-labeled data sets.

3.3 Aggregating Visual Evidence Across Users

After image classification, we could have simply count how many photos have evidence of snowfall or greenery in the content taken in a given time period at a given place. But, in fact, while we are enjoying millions of free and public images on social media websites, we are facing the following problems at the same time. The photographers upload these images completely for their own social and personal

purposes. They may upload a large batch of images without any evidence of our target scene, for example, on a snow day, people can still upload hundreds of indoor photos or close up photos of their dogs or kids or plants in their cleaned backyard. Or simply due to device setting of the camera or smartphone, and sometimes limited GPS accuracy of the photographing device and satellite, it's possible to have a photographer uploading a large number of images with incorrect timestamp or geo-tag. In this case, which is likely to see on social media websites, we count the visual evidence by users instead of by images.

We adopt the likelihood model from **FiXme Note: citewww**, by applying the scene classification model to the training dataset D_1, we figure out for each location during each time period which image is capturing the target scene (snowfall or greenery) and which is not. For the remaining description, we use snowfall as our target scene as an example. First, by combining the user profile of each image, we count the number of users denoting in m providing evidence of snowfall (event u), and the number of users denoting in n uploading photos without snowfall (event $\bar{u}$). We learn a fixed probability of when satellite map shows it's a snow day at a given location, we see a photographer uploading images contain snowfall $p = P(u|snow)$. Similarly, $q = P(s|\overline{snow})$ denotes the probability of when we find a photographer sharing images with snowfall evidence but satellite map shows it's not snowing at that time and location. We assume q is a non-zero probability due to misleading visual contents (photos with very bright wall or a small chance of false positive result from scene classification), and inaccurate geo and temporal tags.

According to **FiXme Note: citewww**, assuming users are taking photos independently, given all observers (photographers) spotting snow or non-snow, we have the posterior probability of snowfall presence at each time and location:

$$\begin{aligned} P(snow|s^m, \bar{s}^n) &= \frac{P(s^m, \bar{s}^n|snow)P(snow)}{P(s^m, \bar{s}^n)} \\ &= \frac{\binom{m+n}{m} p^m (1-p)^n P(snow)}{P(s^m, \bar{s}^n)}, \end{aligned}$$

where we use $P(snow)$ as a general prior probability describing in entire North America, the chance to see a snow day at any time of a year. So we can also derive the similar posterior probability for absence of snowfall,

$$P(\overline{snow}|s^m, \bar{s}^n) = \frac{\binom{m+n}{m} q^m (1-q)^n P(\overline{snow})}{P(s^m, \bar{s}^n)}.$$

We derive the likelihood score by taking the ratio of the snow and non-snow probability. It measures the confidence of snow actually appeared at a given time and place by given all user observations.

$$\frac{P(snow|s^m, \bar{s}^n)}{P(\overline{snow}|s^m, \bar{s}^n)} = \frac{P(snow)}{P(\overline{snow})} \left(\frac{p}{q}\right)^m \left(\frac{1-p}{1-q}\right)^n \quad (1)$$