ABSTRACT

latest writing

==========

workshop

The billions of public photos on online social media sites contain a vast amount of latent visual information about the world. In this paper, we study the feasibility of observing the state of the natural world by recognizing specific types of scenes and objects in large-scale social image collections. More specifically, we study whether we can recreate satellite maps of snowfall by automatically recognizing snowy scenes in geo-tagged, timestamped images from Flickr. Snow recognition turns out to be a surprisingly difficult and under-studied problem, so we test a variety of modern scene recognition techniques on this problem and introduce a large-scale, realistic dataset of images with ground truth annotations. As an additional proof-of-concept, we test the ability of recognition algorithms to detect a particular species of flower, the California Poppy, which could be used to give biologists a new source of data on its geospatial distribution over time.

==========

www

The popularity of social media websites like Flickr and Twitter has created enormous collections of user-generated content online. Latent in these content collections are observations of the world: each photo is a visual snapshot of what the world looked like at a particular point in time and space, for example, while each tweet is a textual expression of the state of a person and his or her environment. Aggregating these observations across millions of social sharing users could lead to new techniques for large-scale monitoring of the state of the world and how it is changing over time. In this paper we step towards that goal, showing that by analyzing the tags and image features of geo-tagged, time-stamped photos we can measure and quantify the occurrence of ecological phenomena including ground snow cover, snow fall and vegetation density. We compare several techniques for dealing with the large degree of noise in the dataset, and show how machine learning can be used to reduce errors caused by misleading tags and ambiguous visual content. We evaluate the accuracy of these techniques by comparing to ground truth data collected both by surface stations and by Earth-observing satellites. Besides the immediate application to ecology, our study gives insight into how to accurately crowd-source other types of information from large, noisy social sharing datasets.

1. INTRODUCTION

latest writing

Research has used Twitter and textual social media to track what is going on with people and the world. But images are better – more information, more faithful, harder to forge, don't have to rely on textual reports. Very little work has used images because they're hard to process automatically. Even the textual work doesn't really consider things at large scale or doesn't measure performance objectively. Here we use images and estimate at continental scale.

We particularly study ecologically related phenomena. Current data is imperfect and ecologists need better data sources, and Flickr could provide that. Current observations are good enough for us to use as ground truth but not perfect. So we can use it to measure our performance but our predictions would still be useful. We choose phenomena that is obvious enough so that scene classification techniques could detect it (as opposed to fine-grained tasks like tracking particular bird species or something).

This is a hard problem. We investigate several techniques to make work better. Metadata (text, timestamps, geotags) eases the problem so we don't have to rely on vision alone. Aggregating information across multiple users means we can make mistakes. Probabilistic confidence interval models the noise explicitly, integrating weak information together. Finally we use deep learning techniques for the vision which are state-of-the-art on classification problems.

==========

workshop

Digital cameras and camera-enabled smartphones are now ubiquitous, with a large fraction of the population taking photos regularly and sharing them online. These millions of people taking pictures form a massive social sensor network that is (in aggregate) observing and capturing the visual world across time and space. Modern phones and cameras record metadata like geo-tags and time-stamps in addition to the images themselves, giving (noisy) calibration information about how this ad-hoc sensor network is arranged. Social media sites like Flickr and Facebook thus contain a large amount of latent visual information about the the world and how it is changing over time.

For instance, many (if not most) outdoor images contain some information about the state of the natural world, such as the weather conditions and the presence or absence of plants and animals (Figure 1). The billions of images on social media sites could be analyzed to recognize these natural objects and phenomena, creating a new source of data to biologists and ecologists. Where are marigolds blooming today, and how is this geospatial distribution different from a year ago? Are honeybees less populous this year than last year? Which day do leaves reach their peak color in each county of the northeastern U.S.? These questions can be addressed to some extent by traditional data collection techniques like satellite instruments, aerial surveys, or longitudal manual surveys of small patches of land, but none of these techniques allows scientists to collect fine-grained data at continental scales: satellites can mon-

Figure 1: Many Flickr images contain evidence about the state of the natural world, including that there is snow on the ground at a particular place and time, that a particular species of bird or animal is present, and that particular species of plants are flowering.

itor huge areas of land but cannot detect fine-grained features like blooming flowers, while manual surveys can collect high-quality and fine-grained data only in a small plot of land. Large-scale analysis of photos on social media sites could provide an entirely new source of data at a fraction of the cost of launching a satellite or hiring teams of biologist observers.

The idea of using crowd-sourced data for science and other purposes is of course not new. Citizen science projects have trained groups of volunteers to recognize and report natural phenomena (like bee counts [**?**], bird sightings [**?**], and snowfall [**?**]) near their homes. Data mining work has shown that social networking sites like Twitter can monitor political opinions [**?**, **?**], predict financial markets [**?**], track the spread of disease [**?**], detect earthquakes [**?**], and monitor weather conditions [**?**]. However, the vast majority of this work has used textual data from micro-blogging sites like Twitter; very few papers have tried to do this with images, despite the fact that images offer evidence that is richer, less ambiguous, and much more difficult to fabricate. This is of course because it is much easier to scan for keywords in Twitter feeds than to automatically recognize semantic content in huge collections of images.

In this paper, we test the feasibility of observing the natural world by recognizing specific types of scenes and objects in large-scale image collections from social media. We consider a well-defined but nevertheless interesting problem: deciding whether there was snow on the ground at a particular place and on a particular day, given the set of publicly-available Flickr photos that were geo-tagged and time-stamped at that place and time. This builds on our early work in Zhang *et al* [**?**] which considered a similar problem, but used only tag information (essentially scanning for photos that had the tag "snow" with some very simple image processing to remove obvious outliers). Here, we explicitly test whether large-scale recognition of the image content itself could be used to do this task. Of course, snow cover can already be monitored through satellites and weather stations (although neither of these data sources is perfect: weather stations are sparse in rural areas and satellites typically cannot estimate snow cover when it is cloudy [**?**]), so this is not a transformative application for ecologists in and of itself. Instead, this is an interesting application for us precisely because fine-grained ground truth is available, so that we can test the accuracy of crowd-sourced observations of the natural world, and judge the feasibility of observing other natural phenomena for which there are no other possible sources of data.

We initially expected snowy scene recognition to be an easy problem, in which just looking for large white regions would work reasonably well. Surprisingly, amongst the hundreds of papers on object and scene classification in the literature, we were surprised to find very few that have explicitly considered detecting snow. A few papers on scene classification include snow-related categories [**?**, **?**, **?**], while a few older papers on natural materials detection [**?**, **?**] consider it along with other categories. We test a variety of recognition techniques on this problem, using a new realistic dataset of several thousand images from Flickr with labeled ground truth. We find that snow detection in consumer images is surprisingly difficult, and we hope this paper and our dataset will help spark interest in this somewhat overlooked vision problem. We also consider an ecology application where reliable data does not exist and Flickr image analysis could be potentially quite valuable: estimating the geo-temporal flowering distribution of the California Poppy.

==========

WWW

The popularity of social networking websites has grown dramatically over the last few years, creating enormous collections of user-generated content online. Photo-sharing sites have become particularly popular: Flickr and Facebook alone have amassed an estimated 100 billion images, with over 100 million new images uploaded every day [**?**]. People use these sites to share photos with family and friends, but in the process they are creating immense public archives of information about the world: each photo is a record of what the world looked like at a particular point in time and space. When combined together, the billions of photos on these sites combined with metadata including timestamps, geo-tags, and captions are a rich untapped source of information about the state of the world and how it is changing over time.

Recent work has studied how to mine passively-collected data from social networking and microblogging websites to make estimates and predictions about world events, including tracking the spread of disease [**?**], monitoring for fires and emergencies [**?**], predicting product adoption rates and election outcomes [**?**], and estimating aggregate public mood [**?**, **?**]. In most of these studies, however, there is either little ground truth available to judge the quality of the estimates and predictions, or the available ground truth is an indirect proxy (e.g. since no aggregate public mood data exists, [**?**] evaluates against opinion polls, while [**?**] compares to stock market indices). While these studies have demonstrated promising results, it is not yet clear when crowd-sourcing data from social media sites can yield reliable estimates, or how to deal with the substantial noise and bias in these datasets. Moreover, these studies have largely focused on textual content and have not taken advantage of the vast amount of visual content online.

In this paper, we study the particular problem of estimating geo-temporal distributions of ecological phenomena using geo-tagged, time-stamped photos from Flickr. Our motivations to study this particular problem are three-fold. First, biological and ecological phenomena frequently appear in images, both because photographers take photos of them purposely (e.g. close-ups of plants and animals) or incidentally (a bird in the background of a family portrait, or the snow in the action shot of children sledding). Second, for the two phenomena we study here, snowfall and vegetation cover,

large-scale (albeit imperfect) ground truth is available in the form of observations from satellites and ground-based weather stations. Thus we can explicitly evaluate the accuracy of various techniques for extracting semantic information from large-scale social media collections.

Third, while ground truth is available for these particular phenomena, for other important ecological phenomena (like the geo-temporal distribution of plants and animals) no such data is available, and social media could help fill this need. In fact, perhaps no community is in greater need of real-time, global-scale information on the state of the world than the scientists who study climate change. Recent work shows that global climate change is impacting a variety of flora and fauna at local, regional and continental scales: for example, species of high-elevation and cold-weather mammals have moved northward, some species of butterflies have become extinct, waterfowl are losing coastal wetland habitats as oceans rise, and certain fish populations are rapidly declining [**?**]. However monitoring these changes is surprisingly difficult: plot-based studies involving direct observation of small patches of land yield high-quality data but are costly and possible only at very small scales, while aerial surveillance gives data over large land areas but cloud cover, forests, atmospheric conditions and mountain shadows can interfere with the observations, and only certain types of ecological information can be collected from the air. To understand how biological phenomena are responding to both landscape changes and global climate change, ecologists need an efficient system for ground-based data collection to give detailed observations across the planet. A new approach for creating ground-level, continental-scale datasets is to use passive data-mining of the huge number of visual observations produced by millions of users worldwide, in the form of digital images uploaded to photo-sharing websites.

Challenges. There are two key challenges to unlocking the ecological information latent in these photo datasets. The first is how to recognize ecological phenomena appearing in photos and how to map these observations to specific places and times. Fortunately, modern photo-sharing sites collect a rich variety of non-visual information about photos, including metadata recorded by the digital camera — exposure settings and timestamps, for example — as well as information generated during social sharing — text tags, comments, and ratings, for example. Many sites also record the geographic coordinates of where on Earth a photo was taken, as reported either by a GPS-enabled camera or smartphone, or input manually by the user. Thus online photos include the ingredients necessary to produce geo-temporal data about the world, including information about content (images, tags and comments), and when (timestamp) and where (geotag) each photo was taken.

The second challenge is how to deal with the biases and noise inherent in online data. People do not photograph the Earth evenly, so there are disproportionate concentrations of activity near cities and tourist attractions. Photo metadata is often noisy or inaccurate; for example, users forget to set the clock on their camera, GPS units fail to find fixes, and users carelessly tag photos. Even photos without such errors might be misleading: the tag "snow" on an image might refer to a snow lily or a snowy owl, while snow appearing in an image might be artificial (as in an indoor zoo exhibit).

This paper. In this paper we study how to mine data from photo-sharing websites to produce crowd-sourced observations of ecological phenomena. As a first step towards the longer-term goal of mining for many types of phenomena, here we study two in particular: ground snow cover and vegetation cover ("green-up") data. Both are critical features for ecologists monitoring the earth's ecosystems. Importantly for our study, these two phenomena have accurate fine-grained ground truth available at a continental scale in the form of observations from aerial instruments like NASA's Terra earth-observing satellites [**?**, **?**] or networks of ground-based observing stations run by the U.S. National Weather Service. This data allows us to evaluate the performance of our crowd-sourced data mining techniques at a very large scale, including thousands of days of data across an entire continent. Using a dataset of nearly 150 million geo-tagged Flickr photos, we study whether this data can potentially be a reliable resource for scientific research. An example comparing ground truth snow cover data with the estimates produced by our Flickr analysis on one particular day (December 21, 2009) is shown in Figure 2. Note that the Flickr analysis is sparse in places with few photographs, while the satellite data is missing in areas with cloud cover, but they agree well in areas where both observations are present. This (and the much more extensive experimental results presented later in the paper) suggests that Flickr analysis may produce useful observations either on its own or as a complement other observational sources.

To summarize, the main contributions of this paper include:

— introducing the novel idea of mining photo-sharing sites for geo-temporal information about ecological phenomena,
— introducing several techniques for deriving crowd-sourced observations from noisy, biased data using both visual and textual tag analysis, and
— evaluating the ability of these techniques to accurately measure these phenomena, using dense large-scale ground truth.

2. RELATED WORK

latest

Accuracy of geo and temporal data on Flickr. Over all the images on Flickr.com, XX% (only a small subset) of them come with geo and temporal data. ?Among these data, XX% of them are accurate enough. ?Flickr provides the geo-location accuracy according to the camera device, and GPS precision. Meanwhile, a lot of works are trying to correct estimate or correct geo-location of Flickr images. [**?**] estimates where images are taken for those missing geo-tags. They are optimizing a graph clustering problem. Attributes in their graph include textual tags, timestamps and vision content. It's inspired by an earlier work [**?**]. [**?**] consider textual meta-data to correct geo tags. They also found for users active on both Flickr and Twitter, the Twitter post at around the same time the images are taken can be a reliable reference to estimate the approximate location.

Timestamp is harder to be accurate due to the same reason of geo-tags and the time-zone problem. In paper [**?**], there is a detail analysis in disagreement of camera time and GPS time. They also estimate a more accurate timestamp when users taking multiple images in a short timespan.

Vegetation classification. [**?**] identifies plant species by leaf images. They focus on accurate leaf segmentation according to color difference of leaf and background, curvature distribution over scale, and nearest neighbor matching.

[**?**] introduces multiple Gist models in scene classification. There

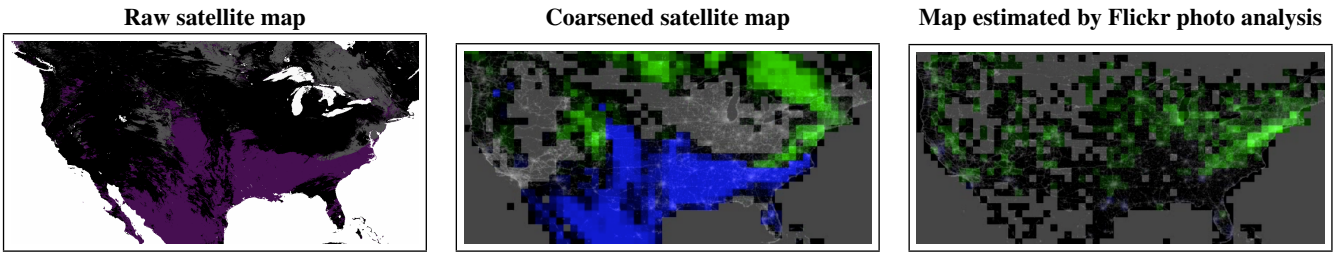


Figure 2: Comparing MODIS satellite snow coverage data for North America on Dec 21, 2009 with estimates produced by analyzing Flickr tags (best viewed on screen in color). *Left:* Original MODIS snow data, where white corresponds with water, black is missing data because of cloud cover, grey indicates snow cover, and purple indicates no significant snow cover. *Middle:* Satellite data coarsened into 1 degree bins, where green indicates snow cover, blue indicates no snow, and grey indicates missing data. *Right:* Estimates produced by the Flickr photo analysis proposed in this paper, where green indicates high probability of snow cover, and grey and black indicate low-confidence areas (with few photos or ambiguous evidence).

happen to be a test set of vegetation shows gist feature works great on vegetation classification.

[**?**] This one may be a little too old It's using man-made mask and neural networks. (could be compared with deep learning)

a lot of works use remote sensing

This paper [**?**] is the closest one to our purpose. They consider color, texture features in images and get good result. But they only test on a very limited dataset where the positive images are either with one tree in the center or full of trees or meadow. This is so not enough when we are working with complex and very large number of public shared images.

==========

workshop

Crowd-sourcing from social media Several recent studies have shown the power of social media for observing the world itself, as a special case of 'social sensing' [**?**]. This work includes using Twitter data to measure collective emotional state [**?**] (which, in turn, has found to be predictive of stock moves [**?**]), predicting product adoption rates and political election outcomes [**?**], and collecting data about earthquakes and other natural disasters [**?**]. Particularly striking examples include Ginsberg *et al* [**?**], who show that geo-temporal properties of web search queries can predict the spread of flu, and Sadilek *et al* [**?**] who show that Twitter feeds can predict when a given person will fall ill.

The specific application we consider here is inferring information about the state of the natural world from social media. Existing work has analyzed textual content, including text tags and Twitter feeds, in order to do this. Hyvarinen and Saltikoff [**?**] use tag search on Flickr to validate metereological satellite observations, although the analysis is done by hand. Zhang *et al* [**?**] take a large collection of geo-tagged and time-stamped Flickr photos and search for snow-related tags to produce estimates of geo-temporal snowfall distributions, and evaluate them against satellite snow maps. Singh *et al* [**?**] visualize geospatial distributions of photos tagged "snow" as an example of their Social Pixels framework, but they study the database theory needed to perform this analysis and do not consider the prediction problem.

Few papers have used actual image content analysis as we do here. Leung and Newsam [**?**] use scene analysis in geo-tagged photos to infer land cover and land use types. Murdock *et al* [**?**] analyze geo-referenced stationary webcam feeds to estimate cloud cover on a day-by-day basis, and then use these estimates to recreate satellite cloud cover maps. Webcams offer a complimentary data source to the social media images we consider here: on one hand, analyzing webcam data is made easier by the fact that the camera is stationary and offers dense temporal resolution; on the other hand, their observations are restricted to where public webcams exist, whereas photos on social media sites offer a potentially much denser spatial sampling of the world.

We note that these applications are related to citizen science projects where volunteers across a wide geographic area send in observations [**?**, **?**, **?**]. These projects often use social media, but require observations to be made explicitly, whereas in our work we "passively" analyze social media feeds generated by untrained and unwitting individuals.

Detecting snow in images We know of only a handful of papers that have explicitly considered snow detection in images. Perhaps the most relevant is the 2003 work of Singhal *et al* [**?**, **?**] which studies this in the context of detecting "materials" like water, grass, sky, etc. They calculate local color and texture features at each pixel, and then compute a probability distribution over the materials at each pixel using a neural network. They partition the image into segments by thresholding these belief values, and assign a label to each segment with a probabilistic framework that considers both the beliefs and simple contextual information like relative location. They find that sky and grass are relatively easy to classify, while snow and water are most difficult. Follow-up work [**?**, **?**] applied more modern techniques like support vector machines. Barnum *et al* [**?**] detect falling snow and rain, a complementary problem to the one we study here of detecting fallen snow.

Papers in the scene recognition literature have considered snowy scenes amongst their scene categories; for instance, Li *et al* [**?**, **?**] mention snow as one possible component of their scene parsing framework, but do not present experimental results. The SUN database of Xiao *et al* [**?**] includes several snow-related classes like "snowfield," "ski slope," "ice shelf," and "mountain snowy," but other categories like "residential neighborhood" sometimes have snow and sometimes do not, such that detecting these scenes alone is not sufficient for our purposes.

Recognizing flowers There are a number of papers on detecting and recognizing flowers in images, although none have specifically considered the California Poppies we study here. Most work on flower classification uses datasets with close-up images of nearly-centered flowers, not the cluttered images typical of Flickr. We use

the work of Nilsback and Zisserman [**?**] as the starting point for our experiments. They perform a binary segmentation step to separate flower from background, represent the foreground with vocabularies of color, shape, and texture features, and then perform recognition using nearest neighbors.

==========

www

A variety of recent work has studied how to apply computational techniques to analyze online social datasets in order to aid research in other disciplines [**?**]. Much of this work has studied questions in sociology and human interaction, such as how friendships form [**?**], how information flows through social networks [**?**], how people move through space [**?**], and how people influence their peers[**?**]. The goal of these projects is not to measure data about the physical world itself, but instead to discover interesting properties of human behavior using social networking sites as a convenient data source.

Crowd-sourced observational data. Other studies have shown the power of social networking sites as a source of observational data about the world itself. Bollen *et al* [**?**] use data from Twitter to try to measure the aggregated emotional state of humanity, computing mood across six dimensions according to a standard psychological test. Intriguingly, they find that these changing mood states correlate well with the Dow Jones Industrial Average, allowing stock market moves to be predicted up to 3 days in advance. However their test dataset is relatively small, consisting of only three weeks of trading data. Like us, Jin *et al* [**?**] use Flickr as a source of data for prediction, but they estimate the adoption rate of consumer photos by monitoring the frequency of tag use over time. They find that the volume of Flickr tags is correlated with with sales of two products, Macs and iPods. They also estimate geo-temporal distributions of these sales over time but do not compare to ground truth, so it is unclear how accurate these estimates are. In contrast, we evaluate our techniques against a large ground truth dataset, where the task is to accurately predict the distribution of a phenomenon (e.g. snow) across an entire continent each day for several years.

Crowd-sourced geo-temporal data. Other work has used online data to predict geo-temporal distributions, but again in domains other than ecology. Perhaps the most striking is the work of Ginsberg *et al* [**?**], who show that by monitoring the geospatial distribution of search engine queries related to flu symptoms, the spread of the H1N1 flu can be estimated several days before the official statistics produced by traditional means. DeLongueville *et al* [**?**] study tweets related to a major fire in France, but their analysis is at a very small scale (a few dozen tweets) and their focus is more on human reactions to the fire as opposed to using these tweets to estimate the fire's position and severity. In perhaps the most related existing work to ours, Singh *et al* [**?**] create geospatial heat maps (dubbed "social pixels") of various tags, including snow and greenery, but their focus is on developing a formal database-style algebra for describing queries on these systems and for creating visualizations. They do not consider how to produce accurate predictions from these visualizations, nor do they compare to any ground truth.

Citizen science. While some volunteer-based biology efforts like the Lost Ladybug Project [**?**] and the Great Sunflower Project [**?**] use social networking sites to organize and recruit volunteer observers, we are not aware of any work that has attempted to passively mine ecological data from social media sites. The visual data in online social networking sites provide a unique resource for tracking biological phenomena: because they are images, this data can be verified in ways that simple text cannot. In addition, the rapidly expanding quantity of online images with geo-spatial and temporal metadata creates a fine-scale record of what is happening across the globe. However, to unlock the latent information in these vast photo collections, we need mining and recognition tools that can efficiently process large numbers of images, and robust statistical models that can handle incomplete and incorrect observations.

3. METHOD

latest writing

Three parts: A. Extracting semantics using tags from individual images, which is what we did in WWW paper. Discuss simple keyword search on e.g. a few snow words, or machine learning to find keyword combinations.

B. Extracting semantics using images from individual images. Here we talk about the traditional features that are covered in ICCV paper, and then the deep learning techniques.

C. Combining evidence together across users. Here we have the simple voting method and the probabilistic confidence score. We can augment with the additional factors that Stefan suggested, including priors based on time of year or geographic location, or other evidence like historical accuracy of specific users.] Also including temporal and spatial smoothing by simply adding to the confidence score priors.

[Note what I am omitting here. My preference would be to ignore any techniques that try to classify bins jointly by aggregating all the tags or visual features of photos together in a bin and then using those features. I would also really like if the final classification is based just on thresholding a confidence score, even if the results are slightly worse than the best we can do. I just think it makes the story more complicated to do otherwise and harder to justify.]

==========

Vision model

For each ecology phenomenon, the most descriptive visual features are extracted from images in training set. We test models built from each feature on testing set. Then these features are normalized and concatenated to build a descriptor of each image. With this combined feature, we learn a vision model on training images. This is the model we use to classify an image as having the phenomenon on it or not.

Deep learning

In this convolutional neural network, we use ImageNet pre-trained model, and further tune it with our hand-labeled snow/vege dataset.

==========

workshop

==========

www

We use a sample of nearly 150 million geo-tagged, timestamped Flickr photos as our source of user-contributed observational data about the world. We collected this data using the public Flickr API, by repeatedly searching for photos within random time periods and geo-spatial regions, until the entire globe and all days between January 1, 2007 and December 31, 2010 had been covered. We applied filters to remove blatantly inaccurate metadata, in particular removing photos with geotag precision less than about city-scale (as reported by Flickr), and photos whose upload timestamp is the same as the EXIF camera timestamp (which usually means that the camera timestamp was missing).

For ground truth we use large-scale data originating from two independent sources: ground-based weather stations, and aerial observations from satellites. For the ground-based observations, we use publicly-available daily snowfall and snow depth observations from the U.S. National Oceanic and Atmospheric Administration (NOAA) Global Climate Observing System Surface Network (GSN) [**?**]. This data provides highly accurate daily data, but only at sites that have surface observing stations. For denser, more global coverage, we also use data from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard NASA's Terra satellite. The satellite is in a polar orbit so that it scans the entire surface of the earth every day. The MODIS instrument measures spectral emissions at various wavelengths, and then post-processing uses these measurements to estimate ground cover. In this paper we use two datasets: the daily snow cover maps [**?**] and the two-week vegetation averages [**?**]. Both of these sets of data including an estimate of the percentage of snow or vegetation ground cover at each point on earth, along with a quality score indicating the confidence in the estimate. Low confidence is caused primarily by cloud cover (which changes the spectral emissions and prevents accurate ground cover from being estimated), but also by technical problems with the satellite. As an example, Figure 2 shows raw satellite snow data from one particular day.

3.1 Estimation techniques

Our goal is to estimate the presence or absence of a given ecological phenomenon (like a species of plant or flower, or a meteorological feature like snow) on a given day and at a given place, using only the geo-tagged, time-stamped photos from Flickr. One way of viewing this problem is that every time a user takes a photo of a phenomenon of interest, they are casting a "vote" that the phenomenon actually occurred in a given geospatial region. We could simply look for tags indicating the presence of a feature – i.e. count the number of photos with the tag "snow" – but sources of noise and bias make this task challenging, including:

— *Sparse sampling:* The geospatial distribution of photos is highly non-uniform. A lack of photos of a phenomenon in a region does not necessarily mean that it was not there.
— *Observer bias:* Social media users are younger and wealthier than average, and most live in North America and Europe.
— *Incorrect, incomplete and misleading tags:* Photographers may use incorrect or ambiguous tags — e.g. the tag "snow" may refer to a snowy owl or interference on a TV screen.
— *Measurement errors:* Geo-tags and timestamps are often incorrect (e.g. because people forget to set their camera clocks).

A statistical test. We introduce a simple probabilistic model and use it to derive a statistical test that can deal with some such sources of noise and bias. The test could be used for estimating the presence of any phenomenon of interest; without loss of generality we use the particular case of snow here, for ease of explanation. Any given photo either contains evidence of snow (event s) or does not contain evidence of snow (event $\bar{s}$). We assume that a given photo taken at a time and place with snow has a fixed probability $P(s|snow)$ of containing evidence of snow; this probability is less than 1.0 because many photos are taken indoors, and outdoor photos might be composed in such a way that no snow is visible. We also assume that photos taken at a time and place without snow have some non-zero probability $P(s|\overline{snow})$ of containing evidence of snow; this incorporates various scenarios including incorrect timestamps or geo-tags and misleading visual evidence (e.g. man-made snow).

Let m be the number of snow photos (event s), and n be the number of non-snow photos (event $\bar{s}$) taken at a place and time of interest. Assuming that each photo is captured independently, we can use Bayes' Law to derive the probability that a given place has snow given its number of snow and non-snow photos,

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

where we write $s^m, \bar{s}^n$ to denote m occurrences of event s and n occurrences of event $\bar{s}$, and where $p = P(s|snow)$ and $P(snow)$ is the prior probability of snow. A similar derivation gives the posterior probability that the bin does not contain snow,

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

where $q = P(s|\overline{snow})$. Taking the ratio between these two posterior probabilities yields a likelihood ratio,

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

This ratio can be thought of as a measure of the confidence that a given time and place actually had snow, given photos from Flickr.

A simple way of classifying a photo into a positive event s or a negative event $\bar{s}$ is to use text tags. We identify a set $\mathcal{S}$ of tags related to a phenomenon of interest. Any photo tagged with at least one tag in $\mathcal{S}$ is declared to be a positive event s, and otherwise it is considered a negative event $\bar{s}$. For the snow detection task, we use the set $\mathcal{S}$={snow, snowy, snowing, snowstorm}, which we selected by hand.

The above derivation assumes that photos are taken independently of one another, which is generally not true in reality. One particular source of dependency is that photos from the same user are highly correlated with one another. To mitigate this problem, instead of counting m and n as numbers of *photos*, we instead let m be the number of *photographers* having at least one photo with evidence of snow, while n is the numbers of photographers who did not upload any photos with evidence of snow.

The probability parameters in the likelihood ratio of equation (1) can be directly estimated from training data and ground truth. For

example, for the snow cover results presented in Section 4.4, the learned parameters are: $p = p(s|snow) = 17.12\%, q = p(s|\overline{snow}) = 0.14\%$. In other words, almost 1 of 5 people at a snowy place take a photo containing snow, whereas about 1 in 700 people take a photo containing evidence of snow at a non-snowy place.

Figure 2 shows a visualization of the likelihood ratio values for the U.S. on one particular day using this simple technique with $\mathcal{S}$={snow, snowy, snowing, snowstorm}. High likelihood ratio values are plotted in green, indicating a high confidence of snow in a geospatial bin, while low values are shown in blue and indicate high confidence of no snow. Black areas indicate a likelihood ratio near 1, showing little conference either way, and grey areas lack data entirely (having no Flickr photos in that bin on that day).

3.2 Learning features automatically

The confidence score in the last section has a number of limitations, including requiring that a set of tags related to the phenomenon of interest be selected by hand. Moreover, it makes no attempt to incorporate visual evidence or negative textual evidence — e.g., that a photo tagged "snowy owl" probably contains a bird and no actual snow. We use machine learning techniques to address these weaknesses, both to automatically identify specific tags and tag combinations that are correlated with the presence of a phenomenon of interest, and to incorporate visual evidence into the prediction techniques.

Learning tags. We consider two learning paradigms. The first is to produce a single exemplar for each bin in time and space consisting of the set of all tags used by all users. For each of these exemplars, the NASA and/or NOAA ground truth data gives a label (snow or non-snow). We then use standard machine learning algorithms like Support Vector Machines and decision trees to identify the most discriminative tags and tag combinations. In the second paradigm, our goal instead is to classify individual *photos* as containing snow or not, and then use these classifier outputs to compute the number of positive and non-positive photos in each bin (i.e., to compute m and n in the likelihood ratio described in the last section).

Learning visual features. We also wish to incorporate visual evidence from the photos themselves. There is decades of work in the computer vision community on object and scene classification (see [**?**] for a recent survey), although most of that work has not considered the large, noisy photo collections we work with here. We tried a number of approaches, and found that a classifier using a simplified version of GIST augmented with color features [**?**, **?**] gave a good trade-off between accuracy and tractability.

Given an image I, we partition the image into a 4×4 grid of 16 equally-sized rectangular regions. In each region we compute the average pixel values in each of the red, green, and blue color planes, and then convert this color triple from sRGB space to the CIELAB color space [**?**]. CIELAB has a number of advantages, including separating greyscale intensity from the color channels and having greater perceptual uniformity (so that Euclidean distances between two CIELAB color triples are approximately proportional to the human perception of difference between the colors). For each region R we also compute the total gradient energy $E(R)$ within the grayscale plane I_g of the image,

$$\begin{aligned} E(R) &= \sum_{(x,y)\in R} ||\nabla I_g(x,y)|| \\ &= \sum_{(x,y)\in R} \sqrt{I_x(x,y)^2 + I_y(x,y)^2}, \end{aligned}$$

where $I_x(x,y)$ and $I_y(x,y)$ are the partial derivatives in the x and y directions evaluated at point (x,y), approximated as,

$$\begin{aligned} I_x(x,y) &= I_g(x+1,y) - I_g(x-1,y), \\ I_y(x,y) &= I_g(x,y+1) - I_g(x,y-1). \end{aligned}$$

For each image we concatenate the gradient energy in each of the 16 bins, followed by the 48 color features (average L, a, and b values for each of the 16 bins), to produce a 64-dimensional feature vector. We then learn a Support Vector Machine (SVM) classifier from a labeled training image set.

4. EXPERIMENTS AND RESULTS

latest writing

Two test cases: snow and vegetation. (Do we have any others?)

For each of these, we learn visual (and maybe text) classifiers for individual images using our hand-labeled datasets. (Also the group datasets Dennis collected over summer?—-?) We also learn parameters of the confidence score using satellite ground truth. (We don't learn visual models or text classifiers from the satellite ground truth.) (—- Visual models is not learnt from satellite but from hand-labeled images. Text classifier is learnt from satellite ground truth to determine snow/vege days or places.)

We test (—- vision model) **on data collected after the training images.** (—- Therefore, we get results of image prediction. And they are clustered according to location and time. Then these images are used to compute confidence score of each place and time period.) **and we compare to satellite ground truth. For each of snow and vegetation, we present two kinds of results:**

** Overall numbers. (the numbers of overall prediction looks pretty good now. I put the results here and we can take it out any time we want)

A. Single place over time: We choose a few places (cities or frequently-photographed places, maybe including a less-photographed place also for fairness) and look at their daily (biweekly for vegetation) estimate over 2-3 years. Plot these versus ground truth. Hopefully point out that the two track well – i.e. that they spot time of leaf changes or major snow storms. Present quantitative results.

B. Single time over place: Choose a few days and show maps of our estimates versus ground truth on these days. Present quantitative results also.

[I would of course be happiest if we also presented results over time and space, i.e. classifying every bin on every day. But the more I think about this, the more I'm not sure it's necessary... the quantitative results are hard to interpret anyway for all there reasons we've seen so far. I think if we did A and B well, we might not need to do this.] (– just a cool illustration or on top right of first page?)

==========

4.1 Test on vegetation cover

Vege hand labeled dataset

crowd sourcing images

Outdoor Greenery
Outdoor non-Greenery
Indoor
Other-modified
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.

Ground truth of vegetation – NDVI index

The ground truth is specified for every 16 days period. And geometrically, north America area is divided into bins of 0.5 by 0.5 degree. This makes the north America area a 120*160 grid map. Only the days and bins with clear enough cloud coverage can be count as useful units. As in [**?**], a green bin must be covered by at least 50% of green vegetation while a non-green bin only has less than 5% of coverage.

Vision model

Vegetation has the characteristic of signature green color, and the leaves of plants have distinctive visual texture. So we employ SIFT feature to analyze the local gradient distribution. And we also extract 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.

Using SIFT histogram, a model is trained and tested with SVM using RBF kernel. The performance is 78.10%.

Color GIST. Similar to color SIFT feature, we also extract GIST feature from each of the RGB color plane.

The performance is 82.58%.

Combine visual features. The combined visual feature is built from concatenating the normalized GIST feature and SIFT histogram. A new model is learnt based on the combined feature. The performance is 85.9%.

Deep learning

(nothing special?)

Confidence score of vege

For an image taken from a place covered by green vegetation at that time, the probability of this image being a green image is 27%. On the other hand, it's only 3% probability to see a green image in a place not covered by enough green vegetation at that time.

measuring the ratio of log likelihood of being a vegetation bin at each time period.

prediction over space and time

We consider north America area has more images uploaded to photo-sharing website, and is also where Ecologists in the US would be interested in the changing color of vegetation. In our work, we define north America as latitude 10° to 70° and longitude -130° to -50°.

==========

workshop

4.2 Snow detection

As noted above, we are aware of very little work that has considered the problem of detecting snow in images: the most relevant work [**?**] considers snow in the context of natural materials classification, but is over 10 years old, uses a small and biased dataset, and does not report classification results. Recent work on scene understanding [**?**] sometimes includes snow-related scenes, but none of this work applies directly to our problem because snow can appear across a range of different scene types. Snow is really an object, not a type of scene, but we are not aware of any work on recognizing snow in the object detection literature.

We thus begin by assembling a large-scale realistic image dataset, and test a variety of modern classification techniques on the problem of snowy scene detection. We use a labeled subset of this dataset to train classifiers and to test their performance, and then apply these classifiers to the problem of generating satellite-like snowfall maps using image analysis on geo-tagged, time-stamped Flickr photos.

Dataset

We collected a large realistic dataset of Flickr images. A subtle but important issue is how to sample these photos. The distribution of geo-tagged Flickr photos is highly non-uniform, with high peaks in population centers and tourist locations. Sampling uniformly at random from Flickr photos produces a dataset that mirrors this highly non-uniform distribution, biasing it towards cities and away from rural areas. Since our eventual goal is to reproduce continental-scale satellite maps, rural areas are very important. An alternative is biased sampling that attempts to select more uniformly over the globe, but has the disadvantage that it no longer reflects the distribution of Flickr photos. Other important considerations include how to find a variety of snowy and non-snowy images, including relatively difficult images that may include wintery scenes with ice but not snow, and how to prevent highly-active Flickr users from disproportionately affecting the datasets.

We strike a compromise on these issues by combining together datasets sampled in different ways. We begin with a collection of about 100 million Flickr photos geo-tagged within North America and collected using the public API (by repeatedly querying at different times and geo-spatial areas, similar to [**?**]). From this set, we considered only photos taken before January 1, 2009 (so that we could use later years for creating a separate test set), and selected: (1) all photos tagged *snow, snowfall, snowstorm,* or *snowy* in English and 10 other common languages (about 500,000 images); (2) all photos tagged *winter* in English and about 10 other languages (about 500,000 images); (3) a random sample of 500,000 images. This yielded about 1.4 million images after removing duplicates. We further sampled from this set in two ways. First, we selected up to 20 random photos from each user, or all photos if a user had less

than 20 photos, giving about 258,000 images. Second, we sampled up to 100 random photos from each $0.1^\circ \times 0.1^\circ$ latitude-longitude bin of the earth (roughly 10km $\times$ 10km at the mid latitudes), yielding about 300,000 images. The combination of these two datasets has about 425,000 images after removing duplicates, creating a diverse and realistic selection of images. We partitioned this dataset into test and training sets on a per-user basis, so that all of any given user's photos are in one set or the other (to reduce the potential for duplicate images appearing in both training and test).

We then presented a subset of these images to humans and collected annotations for each image. We asked people to label the images into one of four categories: (1) contains obvious snow near the camera; (2) contains a trace amount of snow near the camera; (3) contains obvious snow but far away from the camera (e.g. on a mountain peak); and (4) does not contain snow. For our application of reconstructing snowfall maps, we consider (1) and (2) to be positive classes and (3) and (4) to be negative, since snowfall in the distance does not give evidence of snow at the image's geo-tagged location. In total we labeled 10,000 images.

Snow classification

Snow is a somewhat unique visual phenomenon, and we claim that detecting it in images is a unique recognition task. In some cases, snow can be detected by coarse scene recognition: ski slopes or snowy landscapes are distinctive scenes. But snow can appear in any kind of outdoor scene, and is thus like an object. However, unlike most objects that have some distinctive features, snow is simply a white, near-textureless material. (In fact, our informal observation is that humans detect snow not by recognizing its appearance, but by noticing that other expected features of a scene are occluded; in this sense, detecting snow is less about the features that are seen and more about the features that are *not* seen. We leave this as an observation to inspire future work.) We tested a variety of off-the-shelf visual features for classifying whether an image contains fallen snow. We used Support Vector Machines for classification, choosing kernels based on the feature type. Intuitively, color is a very important feature for detecting snow, and thus we focused on features that use color to at least some degree. Our features include:

Color histograms We begin with perhaps the simplest of color features. We build joint histograms in CIELAB space, with 4 bins on the lightness dimension and 14 bins along each of the two color dimensions, for a total of 784 bins. We experimented with other quantizations and found that this arrangement worked best. We encode the histogram as a 784 dimensional feature and use an SVM with a chi-squared distance (as in [**?**]).

Tiny images We subsample images to 16 $\times$ 16 pixels, giving 256 pixels per RGB color plane and yielding a 768 dimensional feature vector. Drastically reducing the image dimensions yields a feature that is less sensitive to exact alignment and more computationally feasible [**?**].

Spatial Moments Tiny images capture coarse color and spatial scene layout information, but much information is discarded during subsampling. As an alternative approach, we convert the image to LUV color space, divide it into 49 blocks using a 7 $\times$ 7 grid, and then compute the mean and variance of each block in each color channel. Intuitively, this is a low-resolution image and a very simple texture feature, respectively. We also compute maximum, minimum, and median value within each cell, so that the final feature vector has 735 dimensions.

Color Local Binary Pattern (LBP) with pyramid pooling LBP represents each 9×9 pixel neighborhood as an 8-bit binary number by thresholding the 8 outer pixels by the value at the center. We build 256-bin histograms over these LBP values, both on the grayscale image and on each RGB color channel [**?**]. We compute these histograms in each cell of a three-level spatial pyramid, with 1 bin at the lowest level, 4 bins in a 2×2 grid at the second level, and 16 bins in a 4×4 grid at the third level. This yields a $(1+4+16)\times4\times256$ = 21504 dimensional feature vector for each image.

GIST We also apply GIST features, which capture coarse texture and scene layout by applying a Gabor filter bank followed by downsampling [**?**], Our variant produces a 1536-dimensional feature vector and operates on color planes. Scaling images to have square aspect ratios before computing GIST improved classification results significantly; [**?**] observed the same effect on a different problem.

We experimented with a number of other features, and found that they did not work well; local features like SIFT and HOG in particular perform poorly, again because snow does not have distinctive visual appearance.

Results

We tested these approaches to detecting snow on our dataset of 10,000 hand-labeled images. We split this set into a training set of 8,000 images and a test set of 2,000 images, sampled to have an equal proportion of snow and non-snow images (so that the accuracy of a random baseline is 50%). Table 1 presents the results. We observe that all of the features perform significantly better than a random baseline. Gist, Color Histograms and Tiny Image all give very similar accuracies, within a half percentage point of 74%. Spatial Moments and LBP features perform slightly better at 76.2% and 77.0%. We also tested a combination of all features by learning a second-level linear SVM on the output of the five SVMs; this combination performed significantly better than any single feature, at 80.5%.

Figure 3 shows classification performance in terms of an ROC curve, as well as a precision-recall curve in which the task is to retrieval photos containing snow. The precision-recall curve shows that at about 20% recall, precision is very near to 100%, while even at 50% recall, precision is close to 90%. This is a nice feature because in many applications, it may not be necessarily to correct classify all images, but instead to find some images that most likely contain a subject of interest. To give a sense for the difficulty and failure modes of our dataset, we show a random sample of correct and incorrect classification results in Figure 5.

Reconstructing satellite snow maps Finally, we tested whether this automated photo classification run on large-scale collections of geotagged, time-stamped social images could be used to approximate snow maps generated by satellites. An advantage of considering the snow recognition task is that ground truth, in the form of daily snow cover maps, is publicly available from NASA and others [**?**].

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%**

Table 1: Performance of different features for snow detection, all using SVMs for classification.

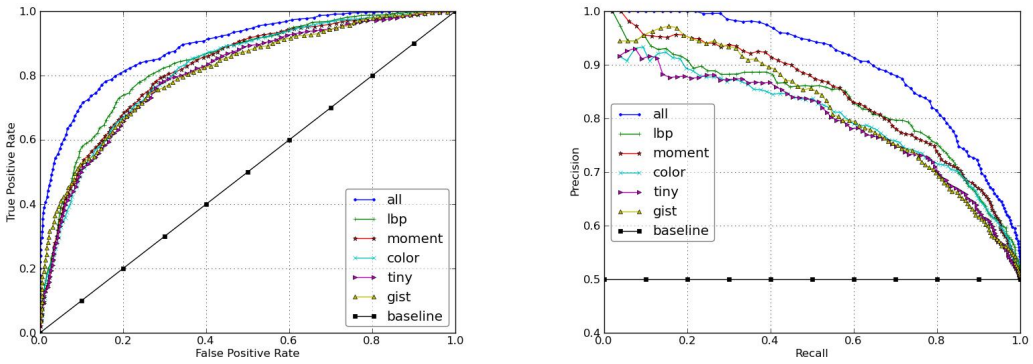


Figure 3: Snow classification results for different features and combinations, in terms of *(left):* ROC curves for the task of classifying snow vs. non-snow images; and *(right):* Precision-Recall curves for the task of retrieving snow images.

This is thus a somewhat artificial task because very good datasets already exist for snow cover, but we use this problem here as a test case of the more general idea of using Flickr to observe nature. (Nevertheless, satellites are also limited because they require the ground to be visible, and thus are not effective when there is cloud cover.)

To test this idea, we downloaded public, geo-tagged, time-stamped Flickr photos taken in North America on three days: March 3, April 6, and December 21 2009 (4422, 5606, and 9906 photos respectively). We ran our combined classifiers on these images. We discretized the image geo-tags into 1 degree by 1 degree bins, and interpreted each snowy image as evidence of snow in that bin and each non-snowy image as evidence against snow in that bin. We combined this evidence together using the simple Bayesian approach proposed by [**?**]. Figure 4 shows the resulting map produced by our automated Flickr analysis, and compares it to the corresponding snow cover map produced by NASA's MODIS instrument [**?**]. We note that the Flickr map is much sparser than the satellite map, especially in sparsely populated areas like northern Canada and the western U.S. On the other hand, the Flickr maps give some observations even when the satellite maps are missing data due to clouds.

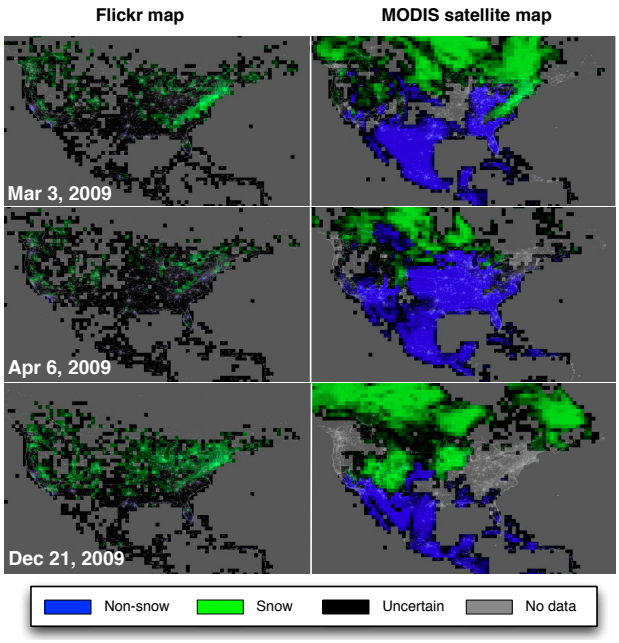


Figure 4: Automatically-generated snow cover maps generated by our Flickr analysis (left), compared with actual satellite maps (right), for three days.

4.3 Detecting California Poppies

We have also studied whether we can apply computer vision analysis of Flickr photos to a problem of interest to biologists: tracking the geo-temporal distribution of flowering plants. Plants and animals will respond as climates change over time, and biologists would like fine-grained, continental scale information about how flowering and migratory patterns are changing. Unlike weather conditions, this data is very difficult to monitor from satellites or aircraft, so biologists currently rely mostly on traditional data collection techniques like longitudinal studies of small plots of land by expert biologists. Analyzing Flickr images could provide an alternative data source for these studies.

As a step in this direction, here we consider one particular class of flower: the California Poppy. We chose this flower both because of its distinct visual appearance (a bright orange) and because it is of interest to biologists because it grows in a relatively small area of the western U.S. and thus may be particularly sensitive to changes in climate.

Dataset

From our collection of about 100 million U.S. Flickr photos, we selected all images tagged "poppy" (about 8100 images). Some of these images are of California Poppies but most are not, since there are other species of poppies and amateur photographers often confuse them with other flowers. We took a random sample of about 2000 images and asked biology students to label them into one of four categories: (1) close-up of a California Poppy; (2) multiple California Poppies (e.g. in a photo of a field perhaps amongst other flowers); (3) no California Poppy; and (4) special cases like drawings of poppies. We discarded images from category (4) and sampled from the remaining dataset to have an equal proportion of the three classes. This gave 150 training images and 450 independent test images. Figure 6 shows a few sample images from our dataset.

Classifying poppies

We used the same features described above in Section 4.2 for classifying snow images, including tiny images, color histograms, color-aware local binary pattern with spatial pyramids, and GIST. For

(a) Random true negatives (non-snow images classified as non-snow)

(b) Random true positives (snow images classified as snow)

(c) Random false negatives (snow images classified as non-snow)

(d) Random false positives (non-snow images classified as snow)

Figure 5: Snow classification results on some random images from our dataset, including (from top) true negatives, true positives, false negatives, and false positives. These results were obtained using the combined classifier that uses all of the image features.

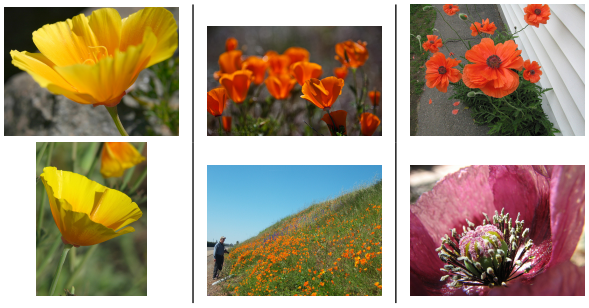


Figure 6: Some images from the three classes in our California Poppy dataset: *(left):* close-ups of true poppies; *(center):* longer-range images of true poppies; and *(right):* images with no poppies.

comparison, we also implemented techniques based on bag-of-words vocabularies of color, texture, and shape features that have been applied to flower recognition in past work [**?**]. In particular:

Color vocabulary We clustered the HSV color values from all images into a 200-word vocabulary, and then represented each image as a histogram over these visual words.

Shape vocabulary We use SIFT features to represent local image "shape." We extracted SIFT features densely (on 25×25 pixel regions, at strides of 20 pixels), and again built a vocabulary using k-means clustering with $k = 200$.

Texture vocabulary We used MR8 features [**?**] to capture local texture information. MR8 applies a filter bank of 8 filters (4 Gaussians and 4 Laplacians of Gaussians) at different scales and orientations, and then characterizes local texture in terms of the maximal filter responses. We again cluster these into a vocabulary with size 200.

We also define a combined feature which incorporates all three of the above features. This feature computes the histogram for each of the three filters, and then concatenates these together after normalizing each vector.

Results

Table 2 shows the performance of the different features on the problem of classifying close-up California Poppy photos versus photos of fields and non-poppies. We observe that the vocabulary-based features work significantly better in combination than separately,

Feature	Accuracy
Random Baseline	33.3%
Shape Vocabulary	45.0%
Texture Vocabulary	48.6%
Color Vocabulary	53.6%
Combination of color, shape and texture	**65.0%**
Tiny Image	58.8%
RGB histogram	61.3%
LBP	68.4%
GIST	68.8%
Spatial pyramid LBP	**70.4%**
Combined	**72.1%**

Table 2: Results for California Poppy classification.

yielding a combined accuracy of 65.0% versus the 33.3% baseline. The LBP, Gist, and LBP features perform better, with the best performance achieved by the combination of features (72.1%). Figure 7 shows ROC and Precision-Recall curves.

==========

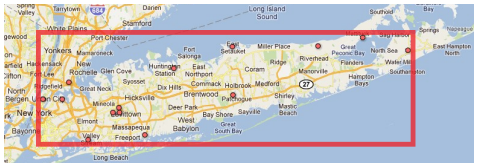
4.4 www – snow coverage

We now turn to presenting experimental results for estimating the geo-temporal distributions of two ecological phenomena: snow and vegetation cover. In addition to the likelihood ratio-based score described in Section 4.4 and machine learning approaches, we also compare to two simpler techniques: *voting,* in which we simply count the number of users that use one of a set S of tags related to the phenomenon of interest at a given time and place, and *percentage,* in which we calculate the ratio of users that use one of the tags in S over the total number of users who took a photo in that place on that day.

Snow prediction in cities

We first test how well the Flickr data can predict snowfall at a local level, and in particular for cities in which high-quality surface-based snowfall observations exist and for which photo density is high. We choose 4 U.S. metropolitan areas, New York City, Boston, Chicago and Philadelphia, and try to predict both daily snow presence as well as the quantity of snowfall. For each city, we define a corresponding geospatial bounding box and select the NOAA ground observation stations in that area. For example, Figure 8 shows the the stations and the bounding box for New York City. We calculate the ground truth daily snow quantity for a city as the average of the valid snowfall values from its stations. We call any day with a non-zero snowfall or snowcover to be a snow day, and any other day to be a non-snow day. Figure 8 also presents some basic statistics for these 4 cities. All of our experiments involve 4 years (1461 days) of data from January 2007 through December 2010; we reserve the first two years for training and validation, and the second two years for testing.

Daily snow classification for 4 cities. Figure 9(a) presents ROC curves for this daily snow versus non-snow classification task on New York City. The figure compares the likelihood ratio confidence score from equation (1) to the baseline approaches (voting and percentage), using the tag set $\mathcal{S}$={snow, snowy, snowing, snowstorm}. The area under the ROC curve (AUC) statistics are 0.929, 0.905, and 0.903 for confidence, percentage, and voting, respectively, and the improvement of the confidence method is statistically significant with $p = 0.0713$ according to the statistical test



	NYC	Chicago	Boston	Philadelphia
Mean active Flickr users / day	65.6	94.9	59.7	43.7
Approx. city area (km^2)	3,712	11,584	11,456	9,472
User density (avg users/unit area)	112.4	52.5	33.5	29.6
Mean daily snow (inches)	0.28	0.82	0.70	0.35
Snow days (snow>0 inches)	185	418	373	280
Number of obs. stations	14	20	41	26

Figure 8: *Top:* New York City geospatial bounding box used to select Flickr photos, and locations of NOAA observation stations. *Bottom:* Statistics about spatial area, photo density, and ground truth for each of the 4 cities.

of [**?**]. The confidence method also outperforms other methods for the other three cities (not shown due to space constraints). ROC curves for all 4 cities using the likelihood scores are shown in Figure 9(b). Chicago has the best performance and Philadelphia has the worst; a possible explanation is that Chicago has the most active Flickr users per day (94.9) while Philadelphia has the least (43.7).

These methods based on presence or absence of tags are simple and very fast, but they have a number of disadvantages, including that the tag set must be manually chosen and that negative correlations between tags and phenomena are not considered. We thus tried training a classifier to learn these relationships automatically. For each day in each city, we produce a single binary feature vector indicating whether or not a given tag was used on that day. We also tried a feature selection step by computing information gain and rejecting features below a threshold, as well as adding the likelihood score from equation (1) as an additional feature. For all experiments we used feature vectors from 2007 and 2008 for training and tested on data from 2009 and 2010, and used a LibLinear classifier with L2-regularized logistic regression [**?**]. Table 3 presents the results, showing that information gain (IG) and confidence scores (Conf) improve the results for all cities, and that the classifier built with both IG and Conf generally outperforms other classifiers, except for Boston. Figure 9(c) shows ROC curves from different classifiers for NYC and Figure 9(d) compares ROC curves for the 4 cities using the classifier using both feature selection and confidence. Note that the machine learning-based techniques substantially outperform the simple likelihood ratio approach (compare Figures 9(b) and (d)).

Predicting snow quantities. In addition to predicting simple presence or absence of a phenomenon, it may be possible to predict the degree or quantity of that phenomenon. Here we try one particular approach, using our observation that the numerical likelihood score of equation (1) is somewhat correlated with depth of snow (R^2=0.2972) — i.e., that people take more photos of more severe storms (see Figure 10). Because snow cover is temporally correlated, we fit a multiple linear regression model in which the confidence scores of the last several days are incorporated. The prediction on day t is then given by,

$$\begin{cases} \sum_{i=0}^{T} \alpha_i \log(conf_{t-i}) + \beta & \text{if } conf_t \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $conf_t$ represents the likelihood ratio from equation (1) on day t, T is the size of the temporal window, and the α and β pa-

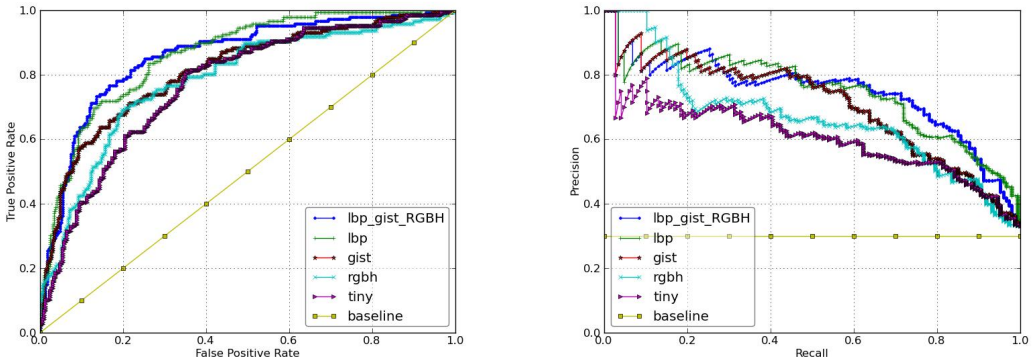
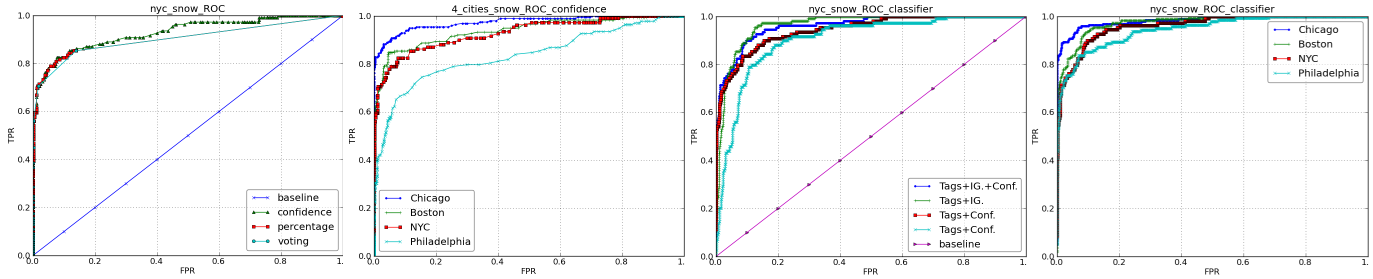


Figure 7: California Poppy classification results for different features, where the goal is to find close-up pictures of California Poppies, in terms of *(left):* ROC curves of classification performance and *(right):* Precision-Recall curves showing retrieval performance.



(a) (b) (c) (d)

Figure 9: ROC curves for binary snow predictions: (a) ROC curves for New York City, comparing likelihood ratio confidence score to voting and percentage approaches, (b) ROC curves for 4 cities using the likelihood scores, (c) ROC curves from SVM classifiers with different features for New York City, and (d) ROC curves for 4 cities using the logistic regression (LibLinear) classifier with tags, information gain and confidence features. (Best viewed in color.)

rameters are learned from the training data. We found that increasing T generally improves performance on the 4 cities, but that no additional improvement occurred with $T > 3$. We can measure the error of our predictions with the root-mean-squared error between

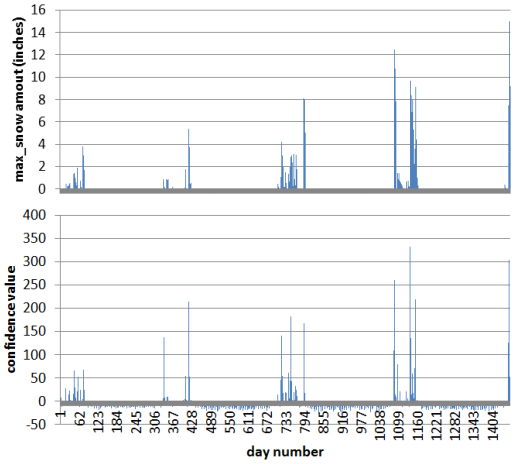


Figure 10: Time series of actual daily snow (top) and score estimated from Flickr (bottom) for New York City, 2007–2010.

Table 3: Daily snow clasification results for a 2 year period (2009–2010) for four major metropolitan areas.

Features	Accuracy	Precision	Recall	F-Measure	Baseline
NYC					
Tags	0.859	0.851	0.859	0.805	0.85
Tags+Conf.	0.926	0.927	0.926	0.917	0.85
Tags+IG	0.91	0.906	0.91	0.898	0.85
Tags+IG+Conf.	**0.93**	**0.93**	**0.93**	**0.923**	**0.85**
Boston					
Tags	0.899	0.897	0.899	0.894	0.756
Tags+Conf.	**0.93**	**0.929**	**0.93**	**0.929**	**0.756**
Tags+IG	0.91	0.911	0.91	0.91	0.756
Tags+IG+Conf.	0.923	0.923	0.923	0.923	0.756
Chicago					
Tags	0.937	0.938	0.937	0.935	0.728
Tags+Conf.	0.949	0.952	0.949	0.948	0.728
Tags+IG	0.938	0.938	0.938	0.938	0.728
Tags+IG+Conf.	**0.953**	**0.954**	**0.953**	**0.953**	**0.728**
Philadelphia					
Tags	0.849	0.851	0.849	0.815	0.805
Tags+Conf.	0.912	0.917	0.912	0.903	0.805
Tags+IG	0.903	0.899	0.903	0.897	0.805
Tags+IG+Conf.	**0.927**	**0.926**	**0.927**	**0.924**	**0.805**

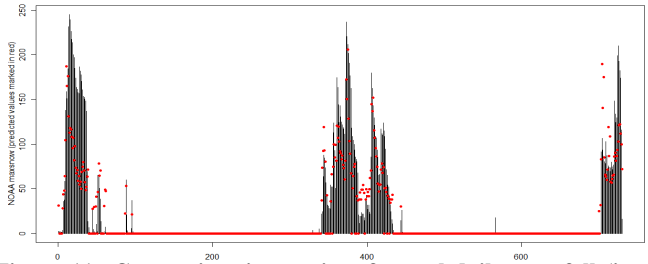


Figure 11: Comparing time series of actual daily snowfall (in mm) for Chicago with estimates using Flickr, for Jan 2009–Dec 2010 and $T = 3$. Red dots show predictions, and vertical bars show actual values.

the time series of our predictions and the actual snow data (following [**?**]). We achieve an RMS error of between about 1 and 1.5 inches across the 4 cities; Philadelphia has the largest error (1.44), followed by Boston (1.26), New York (1.15), and Chicago (1.06). As an example, Figure 11 presents a visual comparison of the prediction time series versus the actual snow time series for Chicago.

An alternative way of evaluating the snow quantity estimates is to view it as a multi-way classification task. We follow an existing snowfall impact scale [**?**] and quantize daily snow quantity into 7 buckets: no snow, 0-1 inches, 1-4 inches, 4-10 inches, 10-20 inches, 20-30 inches, or more than 30 inches. We then build a classifier to predict the snow ranges for the four cities using the numbers of snow and non-snow users. We include the numbers of users from the previous three days as extra features. We use a Naive Bayesian classifier [**?**], which performed best on this task. These multi-way classification results are better than a majority class baseline, with 7-way correct classification rates at 87.5% for Philadelphia, 87.9% for New York, 84.0% for Boston, and 83.7% for Chicago (versus baselines of 80.5%, 85.1%, 75.6%, and 72.9%, respectively).

Continental-scale snow prediction

Predicting snow for individual cities is of limited practical use because accurate meteorological data already exists for these highly populated areas. In this section we ask whether phenomena can be monitored at a continental scale, a task for which existing data sources are less complete and accurate. We use the photo data and ground truth described in Section 4.4, although for the experiments presented in this paper we restrict our dataset to North America (which we defined to be a rectangular region spanning from 10 degrees north, -130 degrees west to 70 degrees north, -50 degrees west). (We did this because Flickr is a dominant photo-sharing site in North America, while other regions have other popular sites — e.g. Fotolog in Latin America and Renren in China.)

The spatial resolution of the NASA satellite ground truth datasets is 0.05 degrees latitude by 0.05 degrees longitude, or about $5 \times 5km^2$ at the equator. (Note that the surface area of these bins is non-uniform because lines of longitude get closer together near the poles.) However, because the number of photos uploaded to Flickr on any particular day and at any given spatial location is relatively low, and because of imprecision in Flickr geo-tags, we produce estimates at a coarser resolution of 1 degree square, or roughly $100 \times 100km^2$. To make the NASA maps comparable, we downsample them to this same resolution by averaging the high confidence observations within the coarser bin. We then threshold the confidence and snow cover percentages to annotate each bin with one of three ground truth labels:

— Snow bin, if confidence is above 90 and coverage above 80,
— Non-snow bin, if confidence is above 90 and coverage is 0,
— Unknown bin, otherwise.

Our goal is to predict whether or not each geospatial bin had snowcover on each day, given the photos from Flickr.

Retrieving snow or non-snow bins. In many real applications, ecologists would be satisfied in finding bins for which the phenomenon is present, rather than actually classifying all bins. It is thus useful to view this problem as a retrieval task, in which the goal is to identify bins likely to contain the phenomenon, or likely not to contain it. We thus turn to evaluating the performance of our estimation techniques using precision-recall curves, where

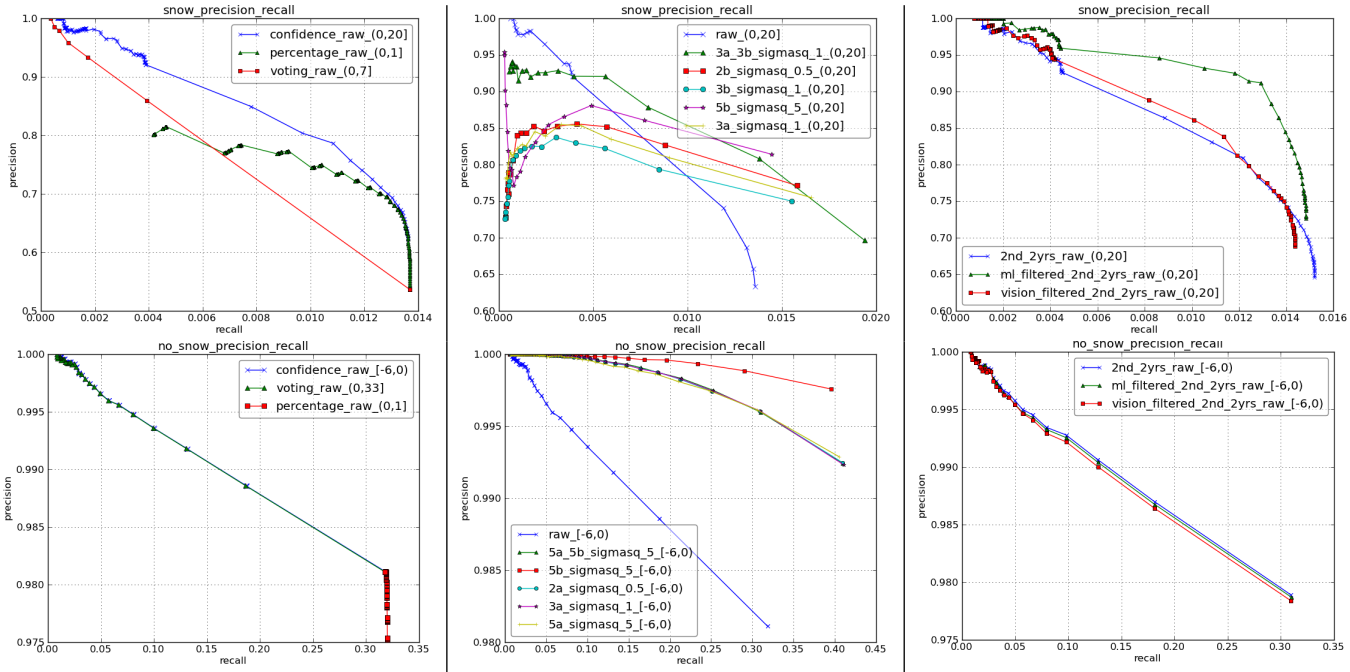
$$\text{precision} = \frac{|R \cap G|}{|R|} \qquad \text{recall} = \frac{|R \cap G|}{|G|},$$

where R is the set of retrieved bins and G is the set of correct bins according to the ground truth. Precision-recall curves are also easier to interpret in situations where the classification baselines are so high, as in our case.

Figure 12(a) shows precision-recall curves for retrieving bins and days containing snow (top) and those not containing snow (bottom). In total, these curves involve classifying about 7 million exemplars (each of which is a single geospatial bin on a single day), of which 11.0% have ground truth. 82.2% of the bins with ground truth are no-snow bins, while snow bins account for 17.8%. We observe that the confidence method performs significantly better than the other two methods for retrieving snow bins, achieving about 98% precision at 0.2% recall, and about 80% precision at 1% recall. For retrieving non-snow bins the three techniques are almost the same, and all three perform better than the random baseline.

While the precisions in these curves are high, the recall values are alarming low. The main reason for this is that large areas of North America, particularly most of Canada and Alaska, have sparse populations resulting in a very limited number of photos uploaded in these areas. We showed in the last section that accurate snow estimates can be inferred for highly populated cities; the low recalls here are because of low photographic density in much of the continent. Restricting to specific subsets significantly increases the density of observations: for example, the average number of photos per bin over our four years of data is nearly ten times larger for the northeast US compared to all of North America (70,398 vs 8,134). The performance is significantly better in these more densely populated areas; for example, in the Northeast US the precision is 96.3% at a recall of 19.5% for snow retrieval, and 99.9% precision at 9.1% recall for non-snow retrieval. Moreover, recall would naturally improve as our dataset grows; our sample of 150 million images is less than 3% of the photos on Flickr, and thus the recall would improve significantly if we had access to the entire dataset.

Temporal smoothing. For many phenomena (including snow), the existence of an event on one day is strongly correlated with its existence on the next day. Thus one way of addressing the sparsity of Flickr photos in some locations is to propagate evidence forward and backward in time. To do this, we apply a Gaussian filter on the Flickr confidence values for each bin in an attempt to achieve better recalls. We vary the degree of smoothing by using Gaussians with different variance values. We tried smoothing with many different parameters, including smoothing both forward and backwards in time, or in only one direction. Figure 12(b) shows curves for several of the best combinations that we found, including the raw confidence score (blue X's), 3 days before and after with variance 1.0 (brown triangles), 2 days before with variance 0.5 (red squares), 3 days before with variance 1.0 (blue circles), 5 days before with variance 5.0 (purple stars), and 3 days after with variance 1.0 (yellow +'s). We find that temporal smoothing three days before and after with variance 1.0 significantly improves performance for both



(a) (b) (c)

Figure 12: Precision and recall curves for retrieving snow (top) and non-snow (bottom) instances, where an instance is a single geospatial bin on a single day, using different techniques: *(a)* comparing the voting, percentage, and statistical confidence estimation techniques, *(b)* comparing different temporal smoothing strategies, *(c)* using classifiers to reject falsely-tagged snow images using visual and textual features.

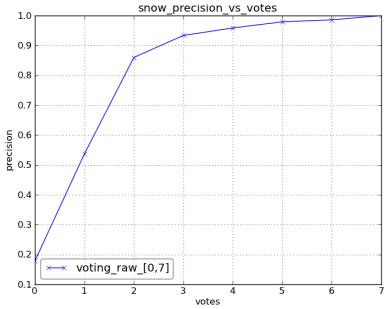


Figure 13: Precision vs number of votes for snow predictions using the voting method.

snow and non-snow retrieval, increasing snow retrieval precision by about 7 percentage points at 1% recall.

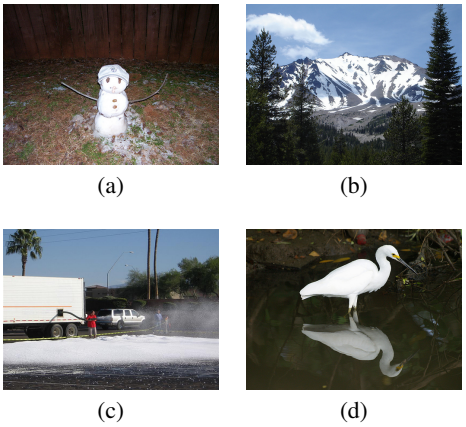
Voting. Voting performs worse than the statistical confidence given by the Bayesian likelihood ratio, but it is an interesting technique to study in more detail because of its simplicity. Voting simply counts the number of users who have annotated at least one photo in a given bin and day with a snow-related tag. Figure 13 plots precision versus the number of votes for snow retrieval. The shape of these curve illustrates why crowd-sourced observations of the world can be reliable, if enough people are involved: as the number of votes for snow increases, it becomes progressively less likely that these independent observations are coincidental, and more likely that they are caused by the presence or absence of an actual phenomenon. It is interesting to notice that when there are 7 or more snow voters, snow prediction precision becomes 100%, while the same is true for non-snow prediction when the number of non-snow voters reaches 33 if there are no snow voters in the bin.

Case study of false positives. To understand the failure modes of estimating attributes about the world from Flickr photos, we performed a case study of false positives — bins and days in which our Flickr mining predicted the presence of snow, but the NASA ground truth indicated that there was no snow cover. In particular, we studied snow false positives at the operating point at which the likelihood ratio method gives a precision of 74.1% and a recall of 1.2% (i.e. when the threshold is 4). At this operating point, 34,323 total predictions are made (each corresponding to a single geospatial bin on a single day), 2,208 of which have valid ground truth. Of these 2,208 bins, 1,636 (74.1%) are correctly classified, while the 572 false positive bins have a total of 1,855 photos tagged with one of the snow terms (despite the fact that they were taken at places and times in which the NASA satellite did not record snow). We manually examined these 1,855 false positive photos and classified them into 5 different classes according to their visual content, as shown in Table 4. Nearly 60% of these photos do actually appear to contain some snow; of these, 33% either show trace amount of snow or snow in the distance (usually on a distant mountain peak), and 8.6% have man-made snow that would not show up on the NASA maps (like in a zoo or ski slope), while only about 16% include a significant amount of natural snow. About 40% of the photos tagged with a snow-related term do not appear to contain any snow at all; these are caused by mis-tagged images or snow-related tags that are used to describe something else (like the interference on a TV screen). Figure 14 shows some sample false positives from each class.

For images that seem to contain natural snow, there are several possible explanations for why the ground truth does not indicate snow cover at that time and place. One is that the satellite passes over at an unknown time of day, so it is possible that snowfall occurred after the satellite's observation was taken. Another cause are photos with incorrect time stamps or geo-locations; we assume that such errors occur frequently, although it is hard to quantify the frequency just by looking at the photos. Other photos clearly contain

Table 4: Taxonomy of manually-labeled false-positive photos (which have at least one snow-related tag despite being taken at a snowless time and place according to the ground truth).

Class	Description	# of photos
little or distant	photos with trace amount of snow or snow in the distance	585 (33.0%)
man made	photos with snow made by humans (e.g. at a ski slope)	152 (8.6%)
no snow	photos without visible snow	737 (41.5%)
snow	photos with significant snow	279 (15.7%)
not sure	other photos	21 (1.2%)



(a) (b) (c) (d)

Figure 14: Sample photos that were not taken at a place and time with snow according to the ground truth, but that were uploaded with a snow-related tag: (a) photo with trace amounts of snow, (b) photo with distant snow, (c) photo with man made snow, and (d) photo with no snow (but with a "snowy egret").

snow, but the amount is so little that it might not be visible from the satellite (e.g. Figure 14(a)), or the snow is so far in the distance that it is in a different geospatial bin (e.g. Figure 14(b)).

There are some cases where the Flickr evidence for snow is overwhelming, but the NASA ground truth does not indicate snow. This could be caused by the timing issue described above, or by satellite resolution and confidence issues. For example, on February 21, 2008, 5 Flickr users reported snowfall in New York. This bin is marked as a no-snow bin in the ground truth because the vast majority of it has zero snow coverage according to the satellite, but there is a small area within the bin that has low confidence (due to cloud cover) and probably corresponds to a snow squall.

Machine learning for tag selection. Many of the above error modes can be addressed by training classifiers on textual tag and visual images features. As discussed in Section 3.1, we are interested in two learning paradigms: the first is to learn combinations of tags that classify geospatial bins well according to the NASA ground truth, while the second task is to reduce false positives by rejecting photos that are tagged with a snow term but do not actually contain snow.

In the first task, we want to learn to classify whether a given bin contains snow on a given day, based on a binary feature vector encoding the set of tags used by all users in that bin on that day. We tried four different classifiers to address this problem: REPTree, a fast decision tree learner which builds a decision tree using information gain and variance and prunes it using reduced-error

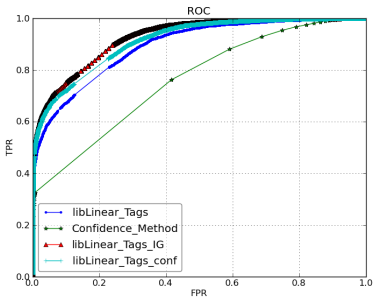


Figure 15: ROC curves for classifying whether a geo-bin has snow on a given day, comparing the LibLinear classifier with various tag features to the confidence method using hand-selected tags.

pruning [**?**], Support Vector Machines (SVMs) [**?**], Discriminative Multinomial Naive Bayes (DMNB) [**?**] and LibLinear classifier with L2-regularized logistic regression [**?**]. To reduce the large number of features (a total of 404324 tags), we compute information gain and keep all features (13442 tags) with information gain greater than zero. Figure 15 presents ROC curves for this task, showing that the learned classifier outperforms the likelihood ratio from equation (1), and that feature selection with information gain and using the confidence ratio as an additional feature all improve performance.

Next we try the second learning paradigm, in which our goal is to examine photos that have a snow-related tag, and use the other tags as well as visual features to decide whether or not they actually contain snow. For example, the classifier might learn that a photo with "snowy" should be discarded if it also contains the tag "egret," since that photo is likely of a bird and not of actual snow. For training these classifiers, we had a human judge evaluate 1,855 images and to annotate them as to whether or not they actually contain evidence of snow.

We used decision trees for this task because it is easy to understand and interpret what features the classifier is using. In initial experimentation, we found that many of the most discriminative features were place names, like "sandiego" or "canada." These geographic tags are understandably strongly correlated with snowfall, but we would like our classifier to base its decisions on the content of an image (because, for example, climate change might cause snowfall in San Diego some day, and we would like our classifier to be able to detect this). To avoid selecting these tags, we first divide North America into four regions (northeast, northwest, southeast, southwest) and get the intersection of the sets of tags used in these four regions. We then use only this set of intersected tags ("InterTags") for building the decision tree. Besides tags, we also tried including the photo's timestamp month as an additional feature.

ROC curves are presented in Figure 16. We see that the time feature helps in improve the results, as does using all tags instead of just the spatially-intersected ones. The baseline (majority class) is 86.3%. It is interesting to examine the top few levels of the trained decision tree, to get a sense for which tags are most discriminative. The top decision node is "summer:" if this tag is present, then the photo is classified as not snow. If summer is not present, then the next few layers look at tags like "mountain," "clouds," "ski," "geese," and "egret."

Machine learning to suppress false positives. Finally, we consider using the photo classifier as a filter while computing the likelihood ratios of Section 3.1, in order to reject photos that are marked with

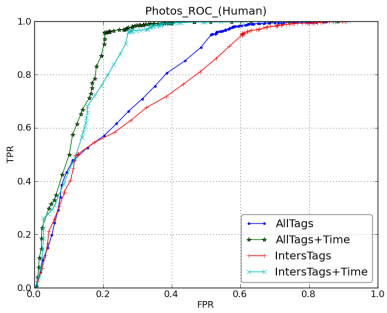


Figure 16: ROC curve for classifying whether photos contain snow, using decision trees with various features: AllTags includes all tags, IntersTags excludes tags corresponding to specific geographic areas, and AllTags+Time and IntersTags+Time include the month of the year as an additional feature.

a snow tag but do not contain snow, using both visual and textual features. For the textual features, we use the decision tree classifier just described. For visual features, we trained an SVM using the GIST-like visual features described in Section 3.1, on the same hand-labeled dataset of about 2,000 images explained above. As with all other experiments, the training and testing sets were kept separate by training on data from 2007-2008 and testing on data from 2009-2010. For the photos in these latter two years, we use our decision tree to try to filter out false positives (photos tagged "snow" but not containing snow), and then re-compute the likelihood ratio confidence score. We find that using a classifier to reject false positives based on tags increased precision by nearly 10 percentage points, as shown in Figure 12(f): at 1% recall, precision increased from about 84% to to about 93% for snow retrieval. For the visual features, we find a significant but more modest improvement, from about 84% to 86% at this level of recall.

4.5 Estimating vegetation cover (may move forward)

Another important measure of the ecological state of the planet is vegetation cover. We perform greenery versus no greenery predictions similarly to snow and no snow predictions using the Flickr confidence threshold method discussed in Section 3.1. As with snow, the ground truth is obtained from down-sampling and thresholding the NASA MODIS greenery data which has the same resolution as the snow cover data with similar coverage and quality (confidence) values. The Flickr greenery confidence values of bins are obtained in a similar way as with snow, except that we use a different set of target tags, including "tree," "trees," "leaf," "leaves", and "grass." One important difference between the NASA greenery and snow datasets is that the greenery data is an average of daily observations spanning 16 days. Thus our goal is to predict the geospatial distribution of greenery for each 16-day period of the year.

We require a bin to have no less than 50% greenery coverage and above middle quality to be considered as a ground truth bin. We report experiments using two different definitions of non-green bins: those having less than 1% coverage, and those having less than 5% coverage. For the 50% and 1% threshold combination, 25.6% of the bins with ground truth are greenery bins, while for the 50% and 5% threshold combination, 15.8% of the bin with ground truth are greenery bins. As shown in Figure 17, both curves outperform a random baseline for greenery prediction, but the estimates are not as accurate as those observed during in snow predictions. There seem to be several reasons for this drop in performance. One is that the boundary between greenery and no greenery seems more vague than the snow/no-snow boundary. Moreover, the greenery ground truth data has a much coarser temporal resolution (16 days). Finally, it's less clear which tags should be used to estimate greenery; using color analysis of the visual content of images may be a better approach, which we leave for future work. We also tried a learned classifier to predict greenery/non-greenery bins based on the set of tags used by all users in each bin and on each day. We used the LibLinear classifier [**?**] because it performed well in case of snow classification. Figure 18 presents the ROC curve for this classification task, showing an equal-error rate of about 91.6%.

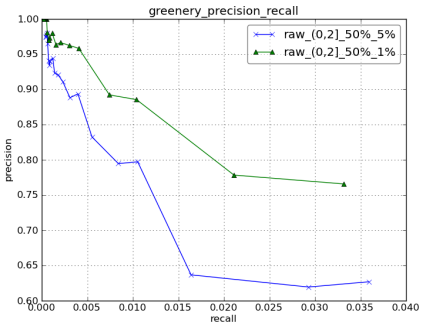


Figure 17: Greenery precision-recall curve using two different ground truth thresholds.

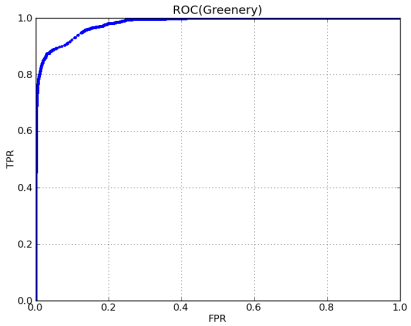


Figure 18: ROC curve for classifying greenery of bins, using tag features and LibLinear classifier.

5. DISCUSSION AND CONCLUSION

latest writing

Big wrap-up, lots of ideas for future work.

==========

workshop

In this paper, we propose using photo-sharing social media sites as a means of observing the state of the natural world, by automatically recognizing specific types of scenes and objects in large-scale social image collections. This work is an initial step towards a long-term goal of monitoring important ecological events and trends through online social media. Our study shows that snowy scene recognition is not nearly as easy a problem as one might expect, when applied to realistic consumer images; our best result using modern vision techniques gives 81% accuracy. Nevertheless, as a proof-of-concept we demonstrated that this recognition accuracy still yields a reasonable map that approximates observations from satellites. We also test recognition algorithms on their ability to recognize a particular species of flower, the California Poppy. In future work, we plan to combine evidence from tags and other metadata with visual features for more accurate estimates, and to develop novel techniques for these challenging recognition problems. More generally, we hope the idea of observing nature through photo-sharing websites will help spark renewed interest in recognizing natural and ecological phenomenon in consumer images.

==========

www

In this paper, we propose using the massive collections of user-generated photos uploaded to social sharing websites as a source of observational evidence about the world, and in particular as a way of estimating the presence of ecological phenomena. As a first step towards this long-term goal, we used a collection of 150 million geo-tagged, timestamped photos from Flickr to estimate snow cover and greenery, and compared these estimates to fine-grained ground truth collected by earth-observing satellites and ground stations. We compared several techniques for performing the estimation from noisy, biased data, including simple voting mechanisms and a Bayesian likelihood ratio. We also tested several possible improvements to these basic methods, including using temporal smoothing and machine learning to improve the accuracy of estimates. We found that while the recall is relatively low due to the sparsity of photos on any given day, the precision can be quite high, suggesting that mining from photo sharing websites could be a reliable source of observational data for ecological and other scientific research. In future work, we plan to study additional features including using more sophisticated computer vision techniques to analyze visual content. Also we plan to study a variety of other ecological phenomena, including those for which high quality ground truth is not available, such as migration patterns of wildlife and the distributions of blooming flowers.