Mining photo-sharing websites to study ecological phenomena

Haipeng Zhang, Mohammed Korayem, David Crandall
School of Informatics and Computing
Indiana University, Bloomington, USA

Gretchen LeBuhn
Department of Biology
San Francisco State University, San Francisco, USA
Social photo sharing websites

- **flickr®**
  - 6+ billion photos

- **facebook**
  - 100+ billion photos

- **Panoramio from Google**
- **Instagram**
  - Fast beautiful photo sharing
- **Fotolog**
- **Kodak Gallery**
- **Picasa™ Web Albums**
- **photobucket**
Spring Blue

Eastern Bluebird, wondering where "his" house is (Answer: Not yet mounted on this pole)!
Need for ecological data

How is nature changing due to global warming?

- **Plot-based studies**: Fine-grained information but only at a few locations, and labor-intensive

- **Aerial surveillance**: Continental-scale information, but only useful for some phenomena

[IPCC2007]
Our paper

• Can we observe nature by mining photo websites?

• We study two phenomena: **snow** and **vegetation cover**
  – Estimate geo-temporal distributions at continental scale, using ~150 million photos from Flickr (via public API)
  – Analyze geo-tags, timestamps, text tags, visual content
  – Evaluate techniques for estimation in crowd-sourced data
  – Compare to data from weather stations and satellites
Related Work

• Crowd-sourced observational data, e.g.:
  – Estimating public mood from Twitter [Bollen11]
  – Predicting product sales from Flickr tags [Jin10]
  – Estimating spread of flu from search queries [Ginsberg09]
  – Monitoring forest fires from Twitter [DeLongueville09]

• Volunteer-based citizen science

  The Great Sunflower Project
  eBird
  The Lost Ladybug Project
Challenges

• Incorrect geotags and timestamps
• Difficult to recognize image content automatically
• Text tags helpful but noisy
  – Some tags are completely incorrect, others are misleading
• Dataset biases
  – Many more photos in cities than rural areas
  – People more likely to take photos of the unusual
• Misleading image data
  – e.g. zoos, ski slopes, synthetic images, etc.
Combining evidence

• Photos by different people are (almost) independent observations, with uncorrelated noise

![Graph showing the probability of actual snow as a function of the number of users tagging a photo with “snow”]
Suppose we’re interested in some object $X$ (e.g. snow)

- Specifically, whether $X$ was present at a given time and place
- Let $s$ denote the event that a given user takes a picture of $X$
- Assume $s$ depends on presence of $X$:

  $P(s \mid X) = \text{probability of taking picture of } X, \text{ given } X \text{ was present}$

  Could be factored into: Probability of seeing $X$, probability of taking photo, probability of uploading to Flickr, ...

  $P(s \mid \bar{X}) = \text{probability of taking picture of } X, \text{ given } X \text{ was not present}$

  Bad timestamps or geotags, misleading image content, ...
A simple model

• Suppose $m$ users took photos of $X$, and $n$ users did not
  – Using Bayes law,

\[
P(X|s^m, \bar{s}^n) = \frac{P(s^m, \bar{s}^n|X)P(X)}{P(s^m, \bar{s}^n)} \quad P(\bar{X}|s^m, \bar{s}^n) = \frac{P(s^m, \bar{s}^n|\bar{X})P(\bar{X})}{P(s^m, \bar{s}^n)}
\]
  – Assuming each user acts independently (conditioned on $X$),

\[
\frac{P(X|s^m, \bar{s}^n)}{P(\bar{X}|s^m, \bar{s}^n)} = \frac{P(X)}{P(\bar{X})} \left( \frac{P(s|X)}{P(s|\bar{X})} \right)^m \left( \frac{1 - P(s|X)}{1 - P(s|\bar{X})} \right)^n
\]
  – High or low ratio means high or low probability of $X$; ratio near 1 means low confidence either way
Snow estimation in cities

• Estimate **daily** snow cover (presence or absence)
  – Predict using Flickr photo tags, compare to ground truth from National Weather Service historical data

Tag set (hand-selected):
{snow, snowy, snowing, snowstorm}

Model parameters (estimated from training data):
P(s | snow) = 17.12%
P(s | no snow) = 0.14%
Learning relevant tags

• Find tags that correlate well with snow cover in GT
  – Feature vector for each day is histogram of number of people that used each tag; labels are snow/no snow from GT
  – Train on 2007-2008 data, test on 2009-2010 data
  – Increases classification accuracies significantly:

![Graphs showing True positive rate vs False positive rate for different cities.](image)
 Continental-scale observation

• Estimate snow cover on each day at each place in North America
  – For each geographic bin of size 1° x 1°
  – Use ground truth data from Terra satellite

[Map showing snow cover and missing data]
Map estimated by Flickr photo analysis

Satellite map (1 degree geo bins)

- Snow cover (green)
- No snow cover (blue)
- Missing data (black/gray)

Dec 21, 2009
Continental-scale estimation

- Predict presence of snow on each day for each geo bin
  - ~35 million total decisions
Visual features

- Color and texture features similar to GIST [Torralba03]
  - Divide image into array of 4x4 cells; in each cell compute mean color value (in CIELab space) and mean gradient energy

Color channels

Image

Gradient magnitude
Visual features

• Color and texture features similar to GIST [Torralba03]
  – Divide image into array of 4x4 cells; in each cell compute
    mean color value (in CIELab space) and mean gradient energy

Image descriptor is concatenation of L, A, B, and G (64 dimensions); then learn SVM classifier
Classification with visual features

- Vision yields modest (~3%) improvement in precision

Correctly classified as non-snow:

Incorrectly classified as snow:
Estimating vegetation cover

• We also estimate vegetation cover (greenery index) on a continental scale
  – Again using ground truth data from Terra satellite
Butterfly
Conclusion

• We propose to observe the natural world through mining public photos from online social sharing sites
  – Hundreds of billions of images available
  – But noise, bias, content extraction are challenges

• We study two phenomena, snow cover and vegetation
  – Using geo-tags, time stamps, text tags, and visual features
  – Use ground truth from satellites to measure estimation accuracy

• Future work
  – More sophisticated computer vision techniques
  – Combine our noisy, sparse data with biologists’ noisy, sparse data
  – Study other phenomena, like migration patterns of wildlife, distributions of blooming flowers, etc.
Thank you!
False positives

- Man-made snow: 9%
- Visible snow, i.e. bad ground truth, timestamps, geotags: 16%
- Trace or distant snow: 33%
- No visible snow, i.e. Incorrect or misleading tags: 42%

(Total of N=1,855 photos)