Towards Extracting Flickr Tag Semantics

Tye Rattenbury[+]

Nathanial Good[*]

Mor Naaman

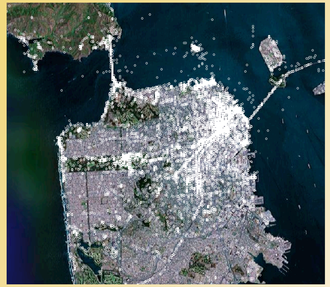
Yahoo! Research Berkeley {tye, ngood, mor}@yahoo-inc.com

[+] Also UC Berkeley, Dept. of Electrical Engineering and Computer Science
[*] Also UC Berkeley, School of Information

Can Flickr help us understand the world? Our system examines the tags associated with geo-tagged and time-stamped photos on Flickr.com. Based on the spatial and temporal distributions of tags (or rather, the spatial and temporal distributions of photos associated with these tags), we try to automatically identify which tags represent places, and which tags represent events.

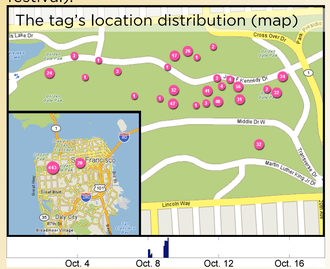
Data

There are over 13,000,000 geo-tagged photos on Flickr. We focused our evaluation on 50,000 photos from the San Francisco Bay Area, and their most frequent tags.



Above: location distribution for all photos in our evaluation.

Below: location & time distributions for tag “Hardly Strictly Bluegrass” (music festival).



The tag's time-based distribution (photos per day)

- 49897 photos
- 803 tags
- ~3.74 tags/photo
- 1015 days
- ~49.16 photos/day

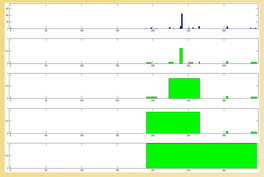
Sample tags (number of photos)

GoldenGatePark	2827
Birthday	1532
Alcatraz	1432
Baseball	1119
GoldenGateBridge	1067
Mission	969
Bay to Breakers	823
parade	722
de Young	714
palaceoffinearts	602
Coit Tower	592
lombardstreet	570
Ocean Beach	560
Fisherman's Wharf	528
hardly strictly bluegrass	525
Embarcadero	519
conservatory of flowers	408
The Future of Web Apps	371
Haight	324
North Beach	322
love parade	298
SOMA	296
eatbrains	264
fleetweek	246
ferrybuilding	228
pier 39	224
unionsquare	219
Apple	193
SF Zoo	190
Exploratorium	152
Cable Car	151

Methods

We analyze two methods inspired by well-known burst-analysis techniques and one novel method: Scale-structure Identification.

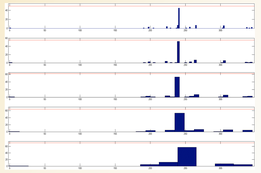
Scale-structure Identification



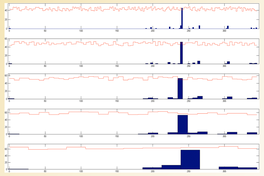
- Perform clustering on the tag's photo distribution using varied inter-cluster distance values
- Compute the cluster entropy for each clustering / inter-cluster distance
- Sum the entropy values and check whether the overall value is smaller than a threshold
- If so, the tag is “structured” according to this distribution

Naïve Scan

- Choose a set of interval sizes
- Segment domain into intervals (one set for each size)
- Count the number of tag usage occurrences in each interval
- For each scale, generate a threshold and check whether any interval has a count above the threshold



Spatial Scan



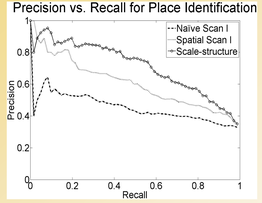
- Choose a set of interval sizes
- Segment domain into intervals (one set for each size)
- Count the number of tag usage occurrences in each interval
- For each scale and each interval, generate a threshold and check whether any interval has a count above the threshold

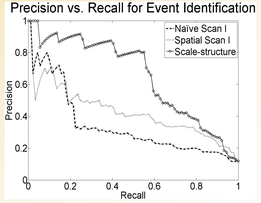
Results

Comparing against ground-truth labels for the tags, Scale-structure Identification outperformed the burst-analysis techniques.

		P-R area	Max F1	Min CE
Naïve Scan	PLACE	0.4455	0.5279	0.2914
Spatial Scan	PLACE	0.6028	0.5907	0.2441
Scale-structure	PLACE	0.7034	0.6655	0.1930
Naïve Scan	EVENT	0.3291	0.3636	0.1009
Spatial Scan	EVENT	0.4130	0.4811	0.1034
Scale-structure	EVENT	0.6420	0.6533	0.0648

- We generated ground truth by hand-labeling tags as `events' or `places'.
- We explored the tradeoff between recall and precision by varying each method's threshold value.





Top ten tags identified as place/event by each method:

Naïve Scan	Alcatraz, *friends*, *event*, PFA, BB, *2006*, China Town, Lombard Street, *trip*, *August*	PLACE
Spatial Scan	GG Park, Alcatraz, Baseball, Giants, HSB, de Young, *event*, SF Giants, PFA, GG Bridge	PLACE
Scale-structure	pet cemetary, Revision3, Ruby Red, Dahlias, *MashPitSF2*, VSHD, Red Devil Lounge, Club Neon, FWA, BH	PLACE
Naïve Scan	BB, eatbrains, eatbrains2006, zombies, *wedding*, zombie, zombiemob2006, byobw, BBW, *Lombard*	EVENT
Spatial Scan	BB, *wedding*, HSB, *event*, byobw, pride, The FWA, BBW, bigwheel, *PFA*	EVENT
Scale-structure	zombiemob, BB 2006, valleyschwag, zombie, zombiemob2006, eatbrains, VSHD, eatbrains2006, zombies, *air race*	EVENT

Future Work

- Apply methods to other geo-referenced data (e.g. Geo-RSS)
- Handle multiple region semantics (e.g. carnival in Brazil vs. Russia)

See Also

- **Full paper at SIGIR 2007**: Tye Rattenbury, Nathan Good, Mor Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. SIGIR 2007, Amsterdam, The Netherlands.