How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections

L. Kennedy, M. Naaman, S. Ahern, R. Nair, T. Rattenbury

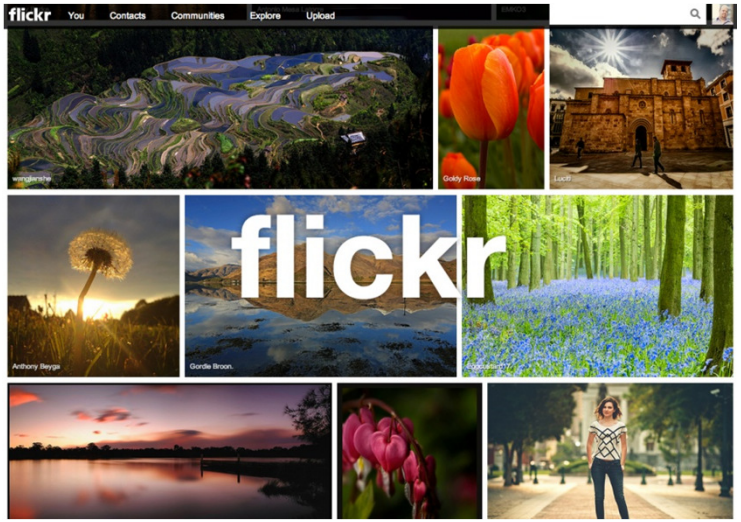
2007

Presenter: Adrien BILLAT

05/02/2014

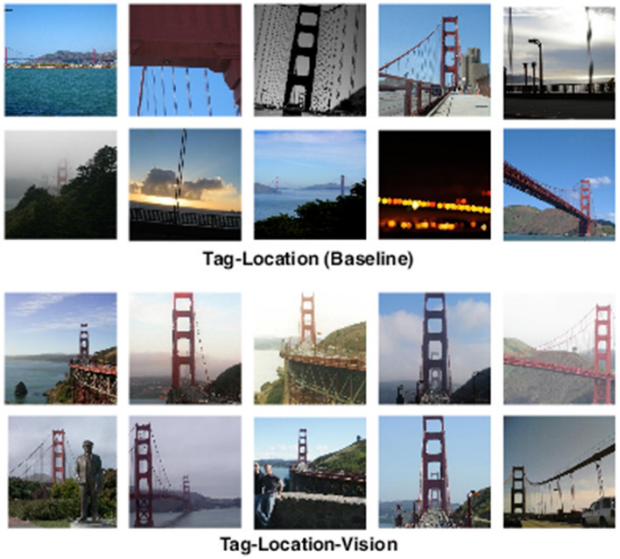
Project Idea

Making sense of the very large and unordered collection of users pictures on Flickr thanks to their associated metadata to improve our access to multimedia resources.



Collection of 20M geo-localized pictures, each of them associated with meta-data:

-location of the photo
-date and time
-user that took it
-set of tags



Collection organized in order to:

-**improve precision** and breadth of retrieval for landmark and **place-based queries**.
-**suggest tags** to un-annotated and geo-referenced pictures.
-generate summaries of sub-collections by selecting **representative photos** for places and identified landmarks.

Computational steps overview

Key contribution: combining tag-based, location-based and image content-based analysis to automatically "**organize**" pictures taken in arbitrary locations of the world.

First step: Identify representative tags for arbitrary locations, by browsing and processing the various tags associated with pictures taken in the area.

Figure 1: Representative tags for San Francisco

Computational steps overview

Second step: Combining the representative tags with their associated location and time in order to identify their semantic. Asses if the tag should be associated with a place or an event by analyzing its spatial or temporal “concentration”.

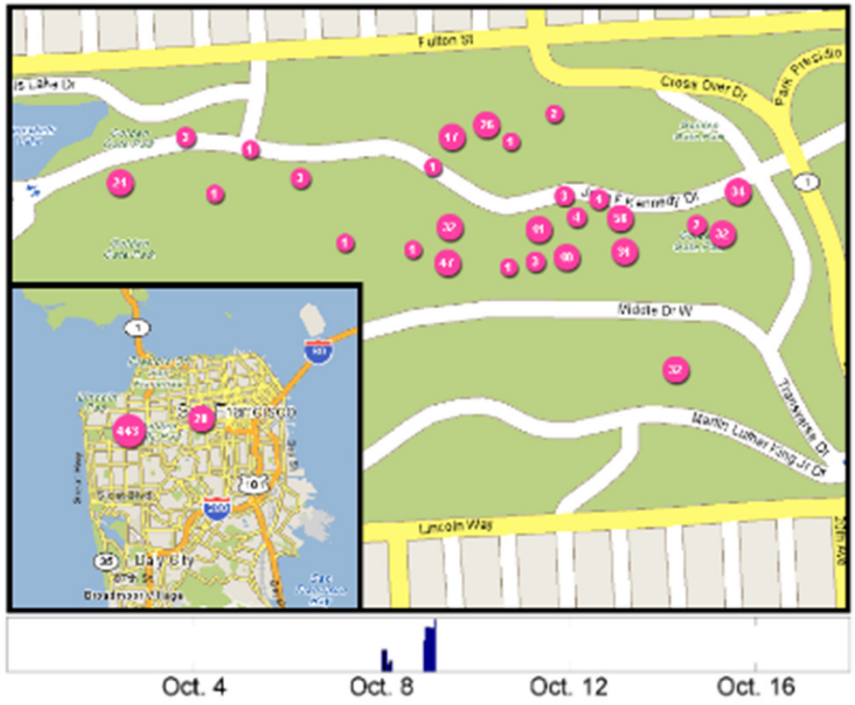


Figure 2: Location (top) and time (bottom) metadata distributions for the tag Hardly Strictly Bluegrass in the San Francisco Bay Area.

Computational steps overview

Third step: Apply visual analysis algorithm on a photos collection deemed as representative of a given landmark/place. They are split into clusters, each of them containing pictures of the landmark from a single viewpoint. Finally, the clusters are sorted by relevance and representative images are extracted to be returned to users queries.

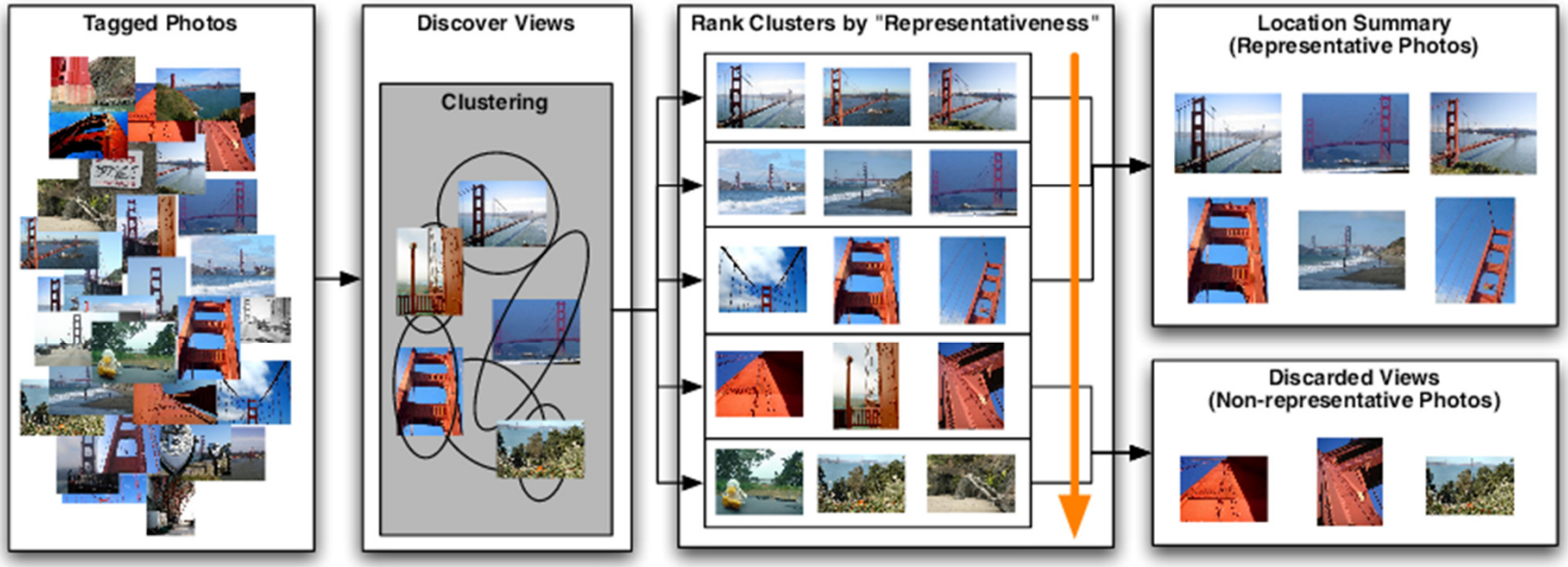


Figure 3: System architecture for choosing representative images from tag/location clusters using visual content.

Mathematical definitions

Basic material for data processing: pictures and tags

Set of pictures $P = \{p\}$ where $p = (\theta p, lp, tp, up) = (\text{picture, location, time, user})$.
Set of tags for picture p $Xp = \{x\}$ where x are arbitrary strings of characters.

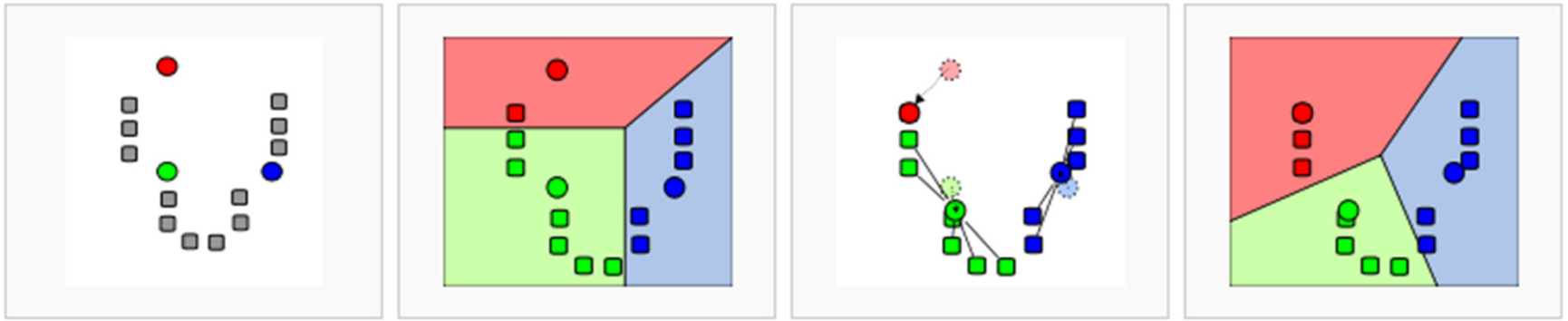
Xs is the set of tags that appear in a subset of pictures $Ps \subset P$.
$Px = \{p \subset P \mid x \subset Xp\}$ is the set of pictures bearing the tag x.

$U = \{up\}$ is the total set of users that took pictures contained in P.
$Ux = \{up \mid p \subset Px\}$ is the set of users using the tag x.

Step 1: representative tags

1) Select a large geographic area (e.g. San Francisco Bay) G and all its photos PG.

2) Geographically cluster the pictures inside G thanks to k-Means algorithm. The numbers of clusters is defined by the number of pictures in the areas (from 3 to 15).



1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

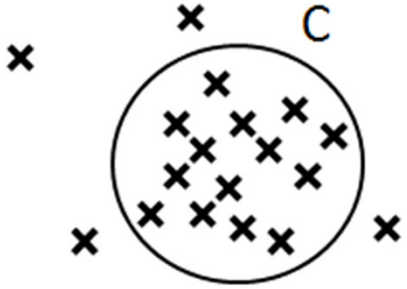
2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3) The centroid of each of the *k* clusters becomes the new mean.

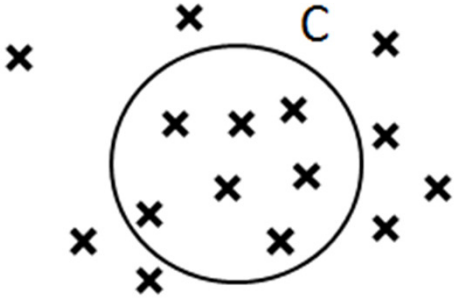
4) Steps 2 and 3 are repeated until convergence has been reached.

Step 1: representative tags

3) Consider each cluster C and its set of tags Xc. Compute the “local relevance” of each tag thanks to its appearance frequency in C and out of C.



relevant distribution



irrelevant distribution

4) Weight the previous results with the variety of users that generated the tag x in C. The more different the photographers are the more representative the tag is. Finally chose the most representative tag.

$$\mathrm{uf} \triangleq \mathbb{U}_{C,x}/\mathbb{U}_C$$

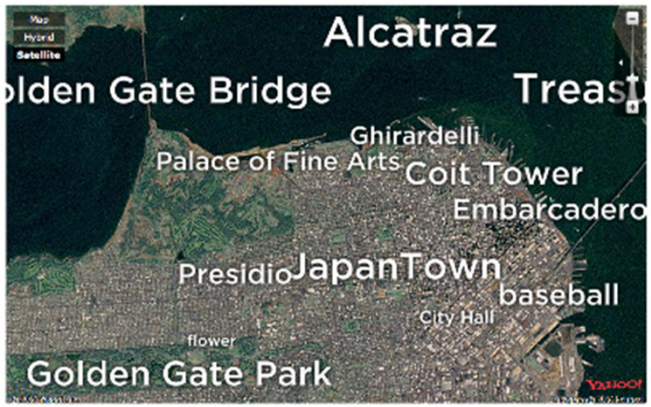


Figure 1: Representative tags for San Francisco

Step 2: tag semantic

1) Once a representative tag x has been computed, identify if its corresponds to a place or an event. Place tags are supposed to exhibit a significant spatial pattern whereas event tags exhibit a significant temporal pattern.

2) Collect Lx = {lp | p C Px} and Tx = {tp | p C Px}. Apply Scale-structure identifications to each collection to measure how similar the distributions are at different scales.
For instance, SSI clusters Lx at different scales, measures how similar this sub-clusters are to a single clusters based on information entropy, sums these entropies to assess how similar Lx is to a single cluster over multiple scale.

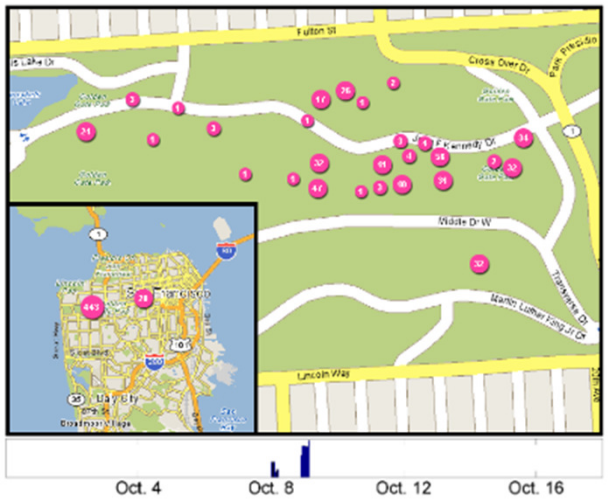


Figure 2: Location (top) and time (bottom) metadata distributions for the tag Hardly Strictly Bluegrass in the San Francisco Bay Area.

Step 3: vision algorithms

Goal: Once a tag has been identified as a landmark, collect and analyze the associated photos to improve the image search. The returned images should:

-be representative of the place.
-feature several viewpoints.
- not contain irrelevant images (excessive zooms, family photos or badly tagged pictures are discarded).

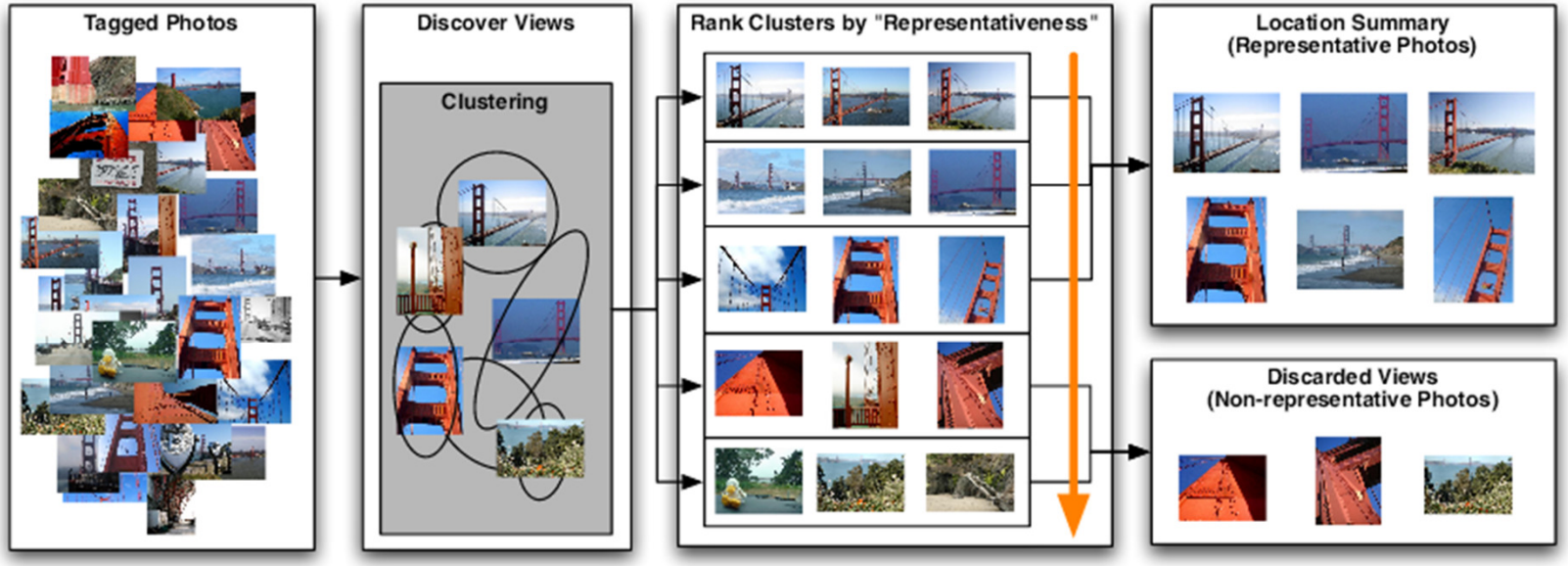


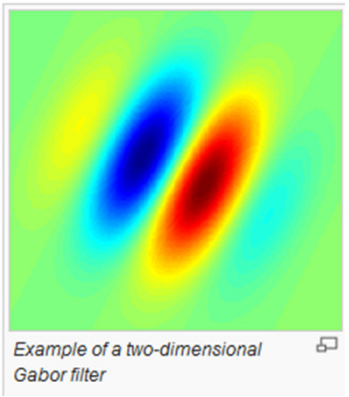
Figure 3: System architecture for choosing representative images from tag/location clusters using visual content.

Step 3: vision algorithms

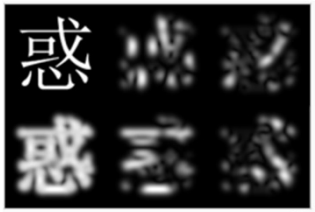
1) Given an identified landmark tag x corresponding to several pictures clusters Cx1, Cx2… We are interested in the set of pictures Px,Cx = Px ∩ {Cx1, Cx2, …}.

2) Extract the following visual features from pictures in Px,Cx:

-Grid color moment i.e. spatial color distribution in LUV space.
-Texture thanks to Gabor transformations of images in several scales and orientation.



Example of a two-dimensional Gabor filter



Demonstration of a Gabor filter applied to Chinese OCR. Four orientations are shown on the right 0°, 45°, 90° and 135°. The original character picture and the superposition of all four orientations are shown on the left.

-Interest points thanks to Scale-Invariant Feature Transform.
About 100 points per image.

Step 3: vision algorithms

3) Cluster the images in Px by viewpoint. Done with k-Means algorithm applied to color and texture data. There should be around 20 images by cluster.

4) Rank the clusters thanks to various indicators computed for each of them:
   -number of users represented.
   -ratio between intra and inter cluster distance (unspecified norm).
   -variability in date.
   -cluster connectivity (using SIFT)

5) Sample these clusters to get Rx, the images representative of tag x.

Evaluation

Example of San Francisco area: 110 000 geo-referenced pictures, 700 clusters.
10 manually selected landmarks are extracted for comparison purpose. The precision is then evaluated manually.

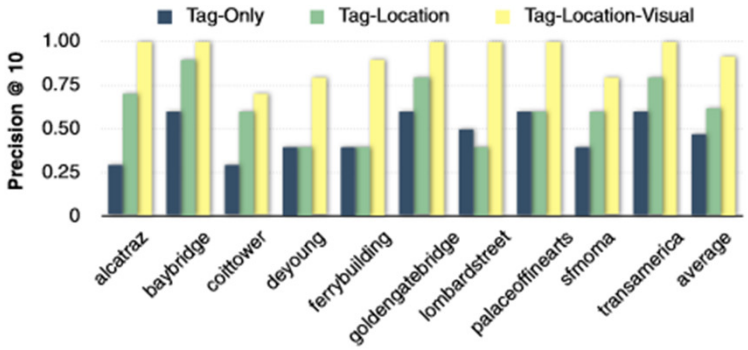


Figure 4: Precision at 10 for representative images selected for locations using various methods.

Precision of top 10 images selected with different methods.

Figure 5: Comparison of recommended representative images resulting from the tag-location filtering and Fixed-size clustering approaches for the Palace of Fine Arts and the Golden Gate Bridge.

Thank you for your attention