Cross-media mining of objects and events from community photo collections

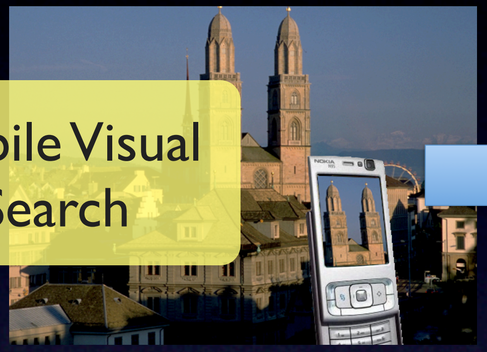
CIAM 2009

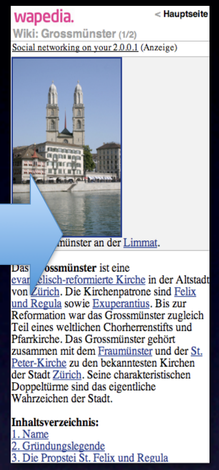
Till Quack
PostDoc BiWi Computer Vision Lab, ETH Zurich, Switzerland
Co-founder kooaba AG

joint work with
Lukas Bossard, Stephan Gammeter, Bastian Leibe, Aleksander Slater and Prof. Luc Van Gool

Motivation

Mobile Visual Search



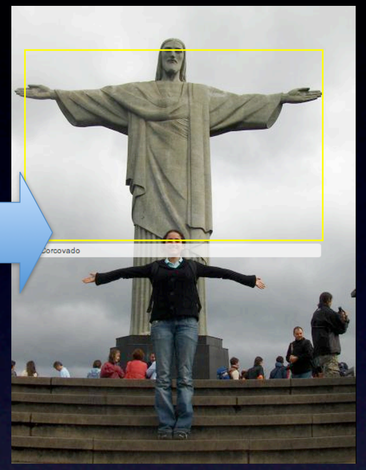


Query Image

Related Content (e.g. Wikipedia)

Auto-annotating holiday snaps





Query Image

Related Content (e.g. Wikipedia, tags)

Cross-media Information Access:

Visual Query

"Text" Response

Powerful visual features
&
Scalable Indexing

SIFT, SURF, MSER — Lowe01, Matas03, Bay06

Visual Vocabularies + Indexing Trees — Sivic03, Nister 06, Philbin07, Chum 08

Cross-media reference database of objects etc.

?

Task

Automatic collection of a reference database for visual recognition from noisy data on the Internet. At large-scale, i.e. efficient.



This talk

1. Cross-media mining pipeline of reference databases for visual recognition

2. Recent extensions to mining pipeline

3. Application prototypes: auto-annotation & Augmented Reality

Cross-media mining of reference databases for visual recognition

Cross-media Mining Pipeline

Collect photos from online communities

↓

Cluster photos

↓

Label and classify clusters (objects vs. events)

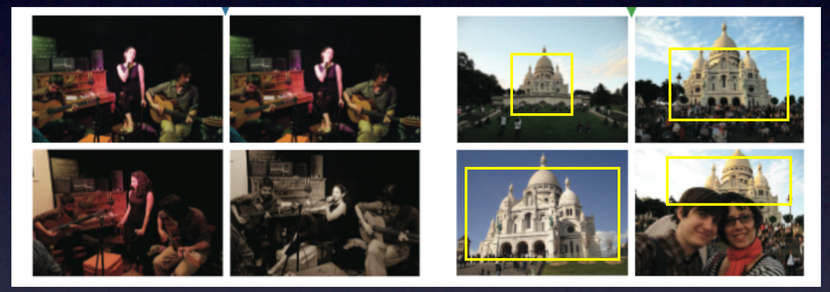
↓

Crawl related content for clusters (cluster description)

↓

Verify description (loop closure)

flickr BETA Panoramio



"event" "object"

http://en.wikipedia.org/wiki/Paris

http://en.wikipedia.org/wiki/Montmartre

http://en.wikipedia.org/wiki/Sacre_coeur

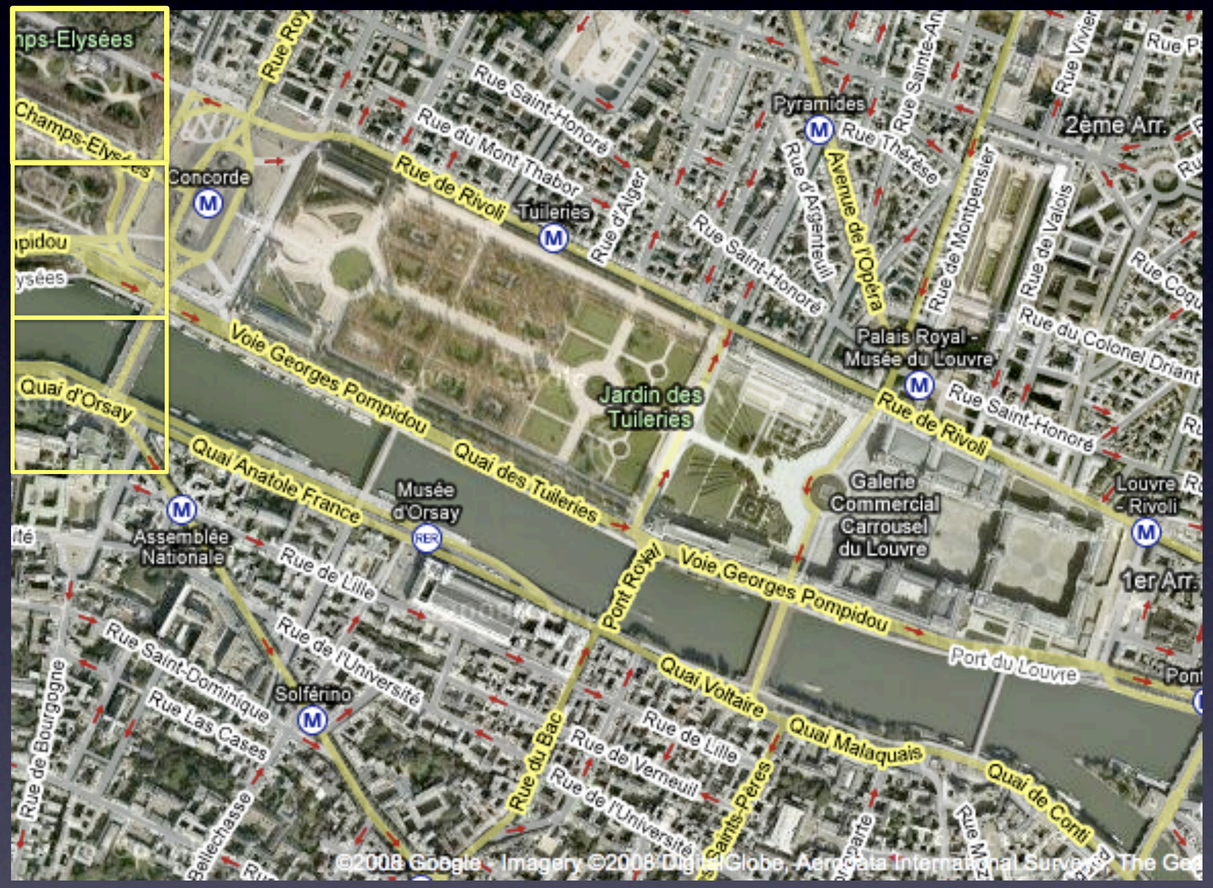
http://en.wikipedia.org/wiki/Sacre_coeur

Collecting Photos

Geo-tagged photos

API call per tile
-width 200m
-step 100m

Retrieve per tile:
- Photos
- Meta-data
  - titles
  - tags
  - time
  - username



http://api.flickr.com/services/rest/?method=flickr.photos.search
&lat=48.861242&lon=2.334423&radius=0.2&radius_units=km

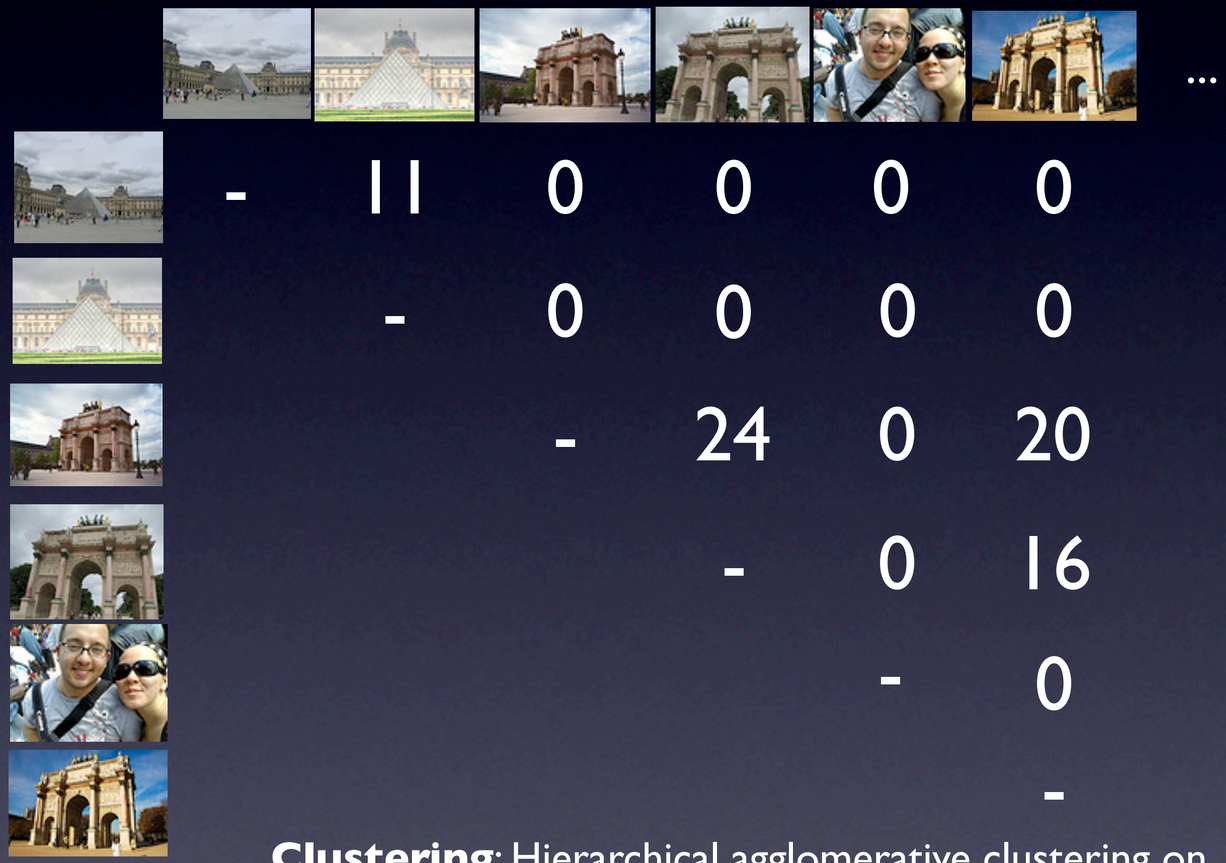
Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Clustering Photos per Tile



- Pair-wise matching of photos per tile
- Gives similarity matrix

Matching per tile scales!
- at most ~10^3 photos per tile
- parallel processing of tiles



	-	11	0	0	0	0
		-	0	0	0	0
			-	24	0	20
				-	0	16
					-	0
						-

Clustering: Hierarchical agglomerative clustering on *distance* matrix

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Clustering: distance measure

Putative SURF feature matches

Inliers I = matches for a Homography H, estimated using RANSAC



$$H\mathbf{x}_n^i = \mathbf{x}_n^j \ , \quad n \in 1 \ldots 4$$

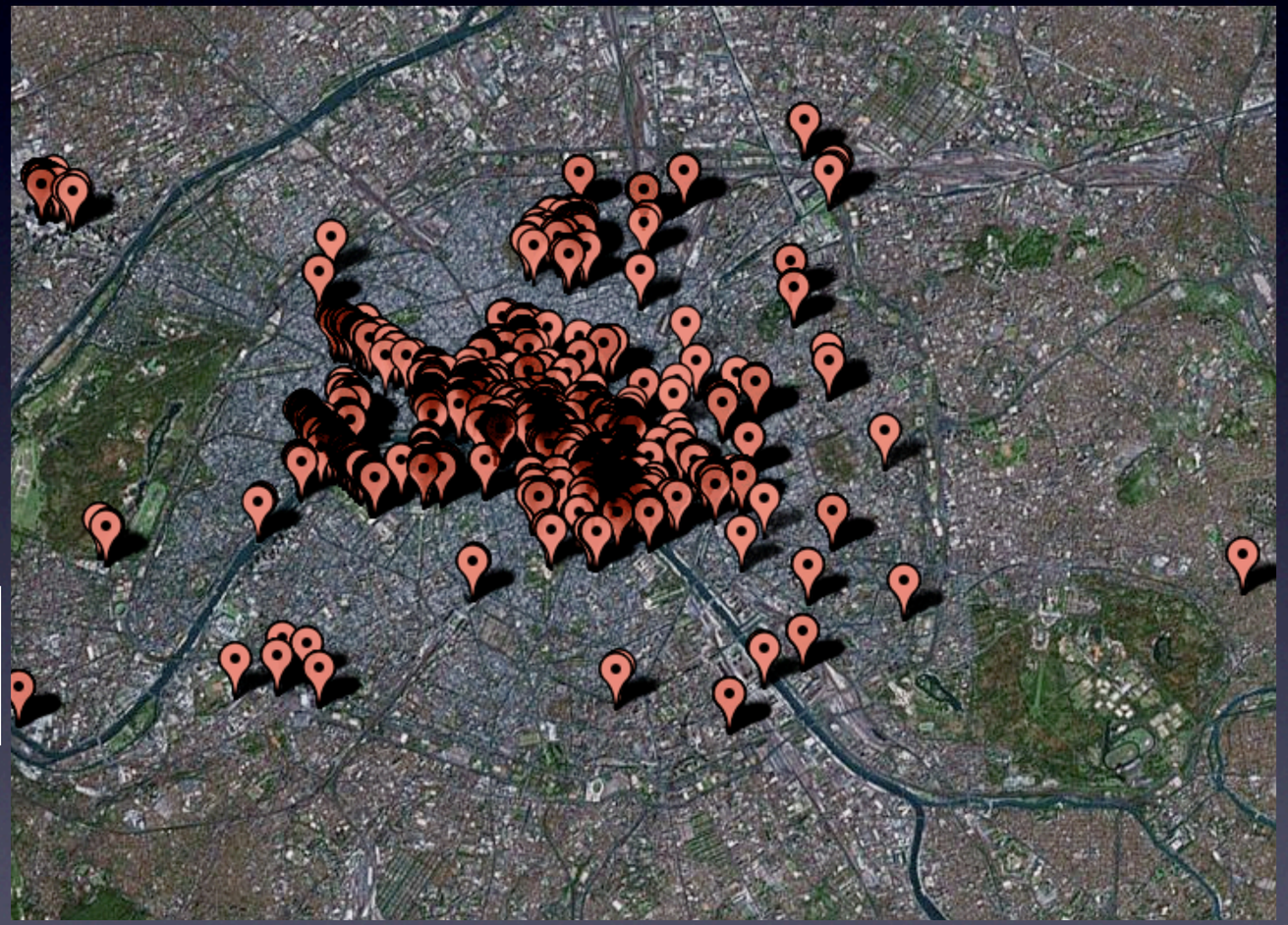
Similarity: number of inliers I_{ij} after RANSAC estimation of Homography H

Distance for clustering

$$d_{ij} = \begin{cases} 1 - \frac{I_{ij}}{I_{max}} & \text{if } I_{ij} \geq 10 \\ \infty & \text{if } I_{ij} < 10 \end{cases}$$

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Sample Clustering Outcome



Matching per tile scales!
- at most ~10^3 photos per tile
- parallel processing of tiles

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Sample Result Pantheon Area



photos in tile: 2250

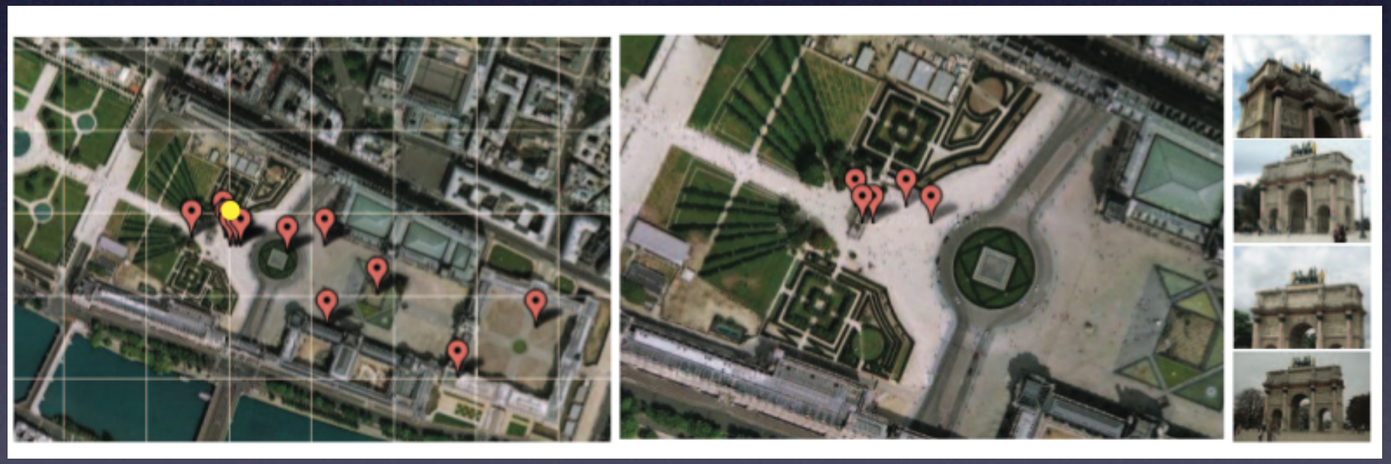
Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Clustering: Single Link vs Complete Link

single link



complete link



All clusters
(mean locations of photos within)

Single cluster
(locations of photos within)

$$\text{single-link:} \quad d_{AB} = \min_{i \in A, j \in B} d_{ij}$$

$$\text{complete-link:} \quad d_{AB} = \max_{i \in A, j \in B} d_{ij}$$

$$\text{average-link:} \quad d_{AB} = \frac{1}{n_i n_j} \sum_{i \in A, j \in B} d_{ij}$$

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Cluster Classification



“event” “object”

- Based on meta-data (cross-medial)
  - **Time**stamps: Number of days pictures were taken at: |D|
  - **User** data: number of users who took photos: |U|
  - Number of photos: |N|
- Simple ID3 tree (stump) on two features

$$f_1 = |D|$$

$$f_2 = \frac{|U|}{|N|}$$

Cross-media content mining

Clustering

VISUAL Features



museum, louvre, carrousel, triomphe , arc du triomphe , img_7312, paris...

Location-specific stoplist

Frequent Itemset Mining on associated tags (Agrawal 93)

TEXT

museum
museum louvre
carrousel
carrousel triomphe
carrousel triomphe arc
carrousel triomphe arc du
carrousel triomphe du
carrousel arc
carrousel arc du
carrousel du
triomphe
triomphe arc
triomphe arc du
triomphe du
arc
arc du

Google

TEXT Retrieval

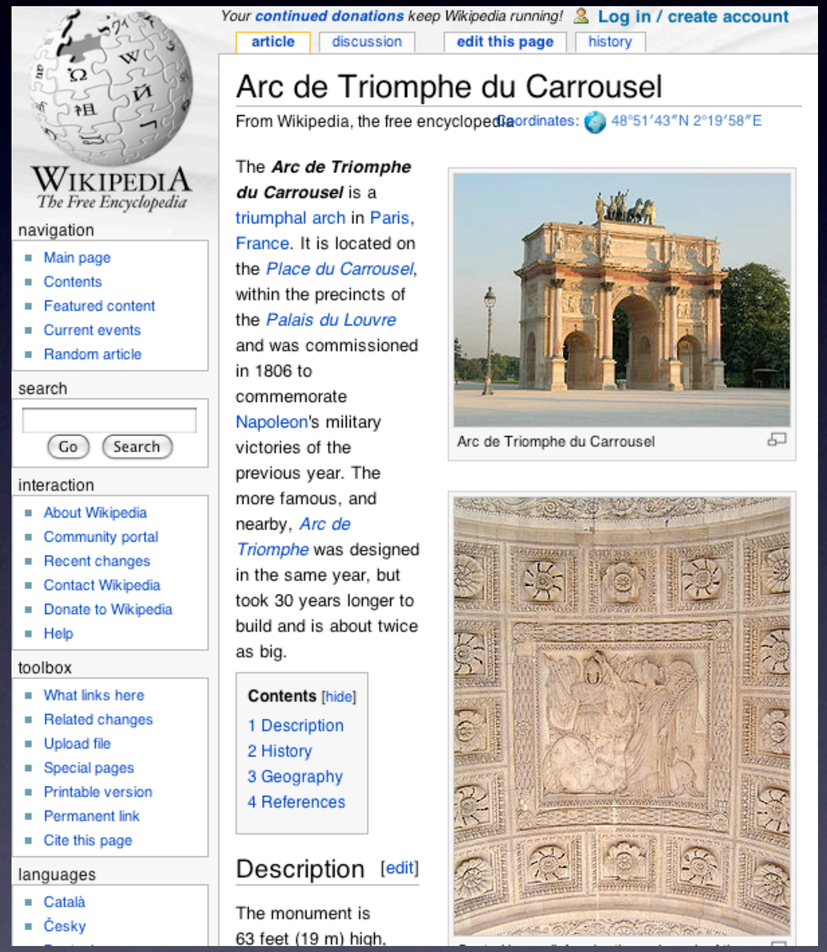
http://en.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca.
http://en.wikipedia.org/wiki/Axe_historique
http://fr.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://en.wikipedia.org/wiki/Quadriga
http://hu.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://de.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://nl.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://it.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://en.wikipedia.org/wiki/Triumphal_arch

Frequent Itemsets serve both as cluster labels and queries to Google

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Cross-media content verification

http://en.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca.
http://en.wikipedia.org/wiki/Axe_historique
http://fr.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://en.wikipedia.org/wiki/Quadriga
http://hu.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://de.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://nl.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://it.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://en.wikipedia.org/wiki/Triumphal_arch



Parse article for images and match back to cluster

Visual Features



Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Wikipedia Matching Examples





(2) http://en.wikipedia.org/wiki/Moulin_Rouge
66 Elements, 39 users, 50 days. Precision: 100%




(7) http://en.wikipedia.org/wiki/Panth%C3%A9on,_Paris
48 elements, 31 users, 37 days. Precision: 98%




(3) http://en.wikipedia.org/wiki/Temple_of_Apollo_Sosianus
33 elements, 22 users, 33 days. Precision: 98.4%




(8) http://en.wikipedia.org/wiki/Notre_Dame_de_Paris
588 elements, 287 users, 334 days. Precision: 100%

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Wikipedia Matching Examples




(4) http://en.wikipedia.org/wiki/Colosseum
http://no.wikipedia.org/wiki/Colosseum
http://sv.wikipedia.org/wiki/Colosseum
582 elements, 190 users, 252 days. Precision: 100%




(5) See (4), matchted to the same cluster.




(9) http://en.wikipedia.org/wiki/Tour_Montparnasse
40 elements, 10 users, 11 days. Precision: 100%




(10) http://en.wikipedia.org/wiki/Campo_dei_Miracoli
http://it.wikipedia.org/wiki/Battistero_di_Pisa
33 elements, 24 users, 21 days. Precision: 94%

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Wikipedia Matching Examples

Lucky shots





(6) http://en.wikipedia.org/wiki/Viktualienmarkt
http://en.wikipedia.org/wiki/Maypole
8 elements, 7 users, 7 days. Precison: 100%.





(9) http://en.wikipedia.org/wiki/Zachary_Quinto
7 elements, 1 users, 1 days . Precision: 100.

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Summary of media-types used

GPS + TIME + USER + VISUAL Features



museum, louvre, carrousel, triomphe , arc du triomphe , img_7312, paris...

TEXT Mining

museum
museum louvre
carrousel
carrousel triomphe
carrousel triomphe arc
carrousel triomphe arc du
carrousel triomphe du
carrousel arc
carrousel arc du
carrousel du
triomphe
triomphe arc
triomphe arc du
triomphe du
arc
arc du

VISUAL Features

TEXT Retrieval



http://en.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca.
http://en.wikipedia.org/wiki/Axe_historique
http://fr.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://en.wikipedia.org/wiki/Quadriga
http://hu.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://de.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://nl.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://it.wikipedia.org/wiki/Arc_de_Triomphe_du_Ca...
http://en.wikipedia.org/wiki/Triumphal_arch

HYPERTEXT analysis

Collect photos ▶ Cluster photos ▶ Label and classify ▶ Crawl related content ▶ Verify description

Experiments

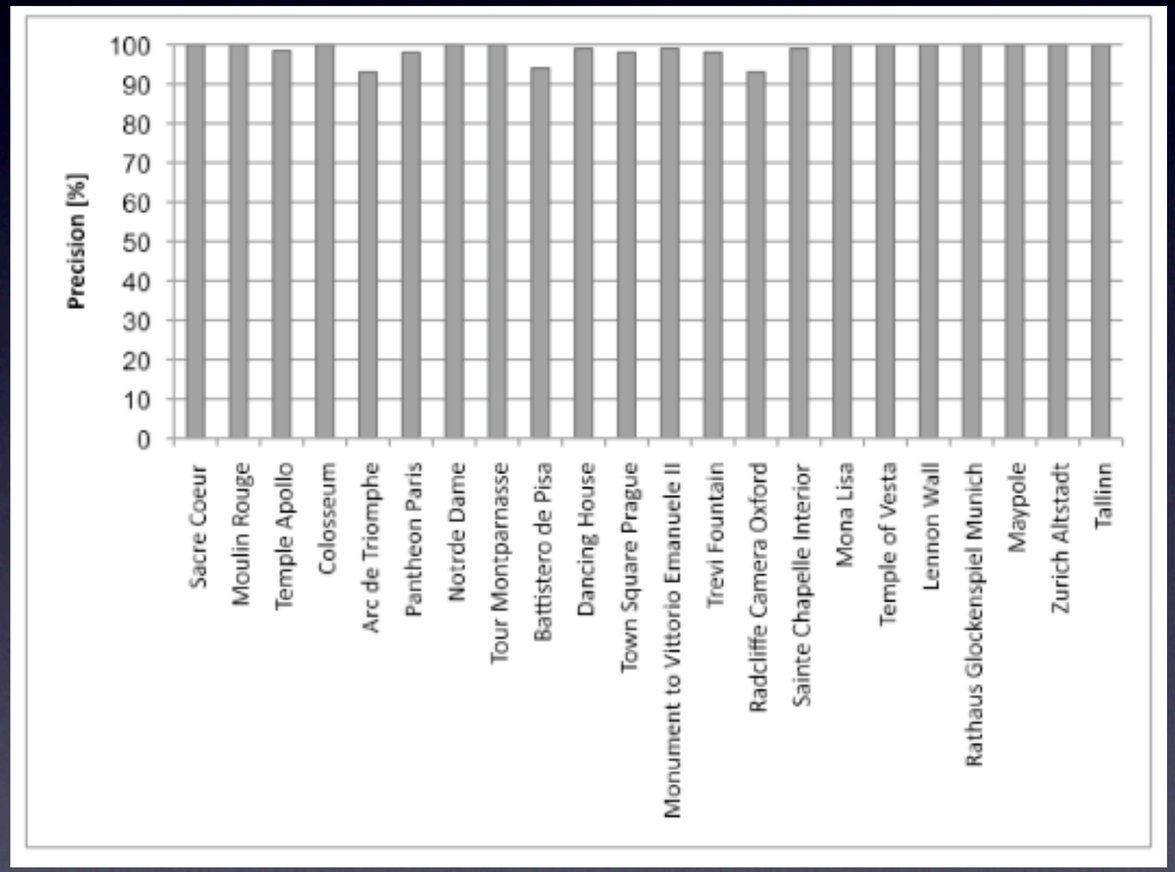
Name	# tiles	#photos	area (km^2)
Munich	18'228	24'069	184.99
Oxford	2'112	7'431	22.05
Paris	12'532	87'452	127.57
Pisa	723	1'950	7.78
Prague	11'110	28'872	113.22
Rome	14'397	48'750	146.38
St. Petersburg	3'400	2'573	35.18
Tallinn	890	1'350	9.51
Venice	449	7'708	4.92
Zurich	5'663	12'602	58.15
Total	**69'504**	**222'757**	**709.74**

# Images	222'757
Size Metadata	1.1 GB
Size Features	111 GB
# Images assigned to clusters	73'236
# Similarities computed	217'330'144
# Similarities > 0	751'457

52'000 empty tiles, average 10 images, max 3750 images.

(State early 2008. by today ~ 4 million images processed)

Clustering Precision



Manual count for select clusters

Precision:
correct images in cluster
High precision in general

Recall:
difficult to assess, but using vision can improve recall. (vs. text based methods)

Example “indoor” / ”outdoor”




Classification Accuracy

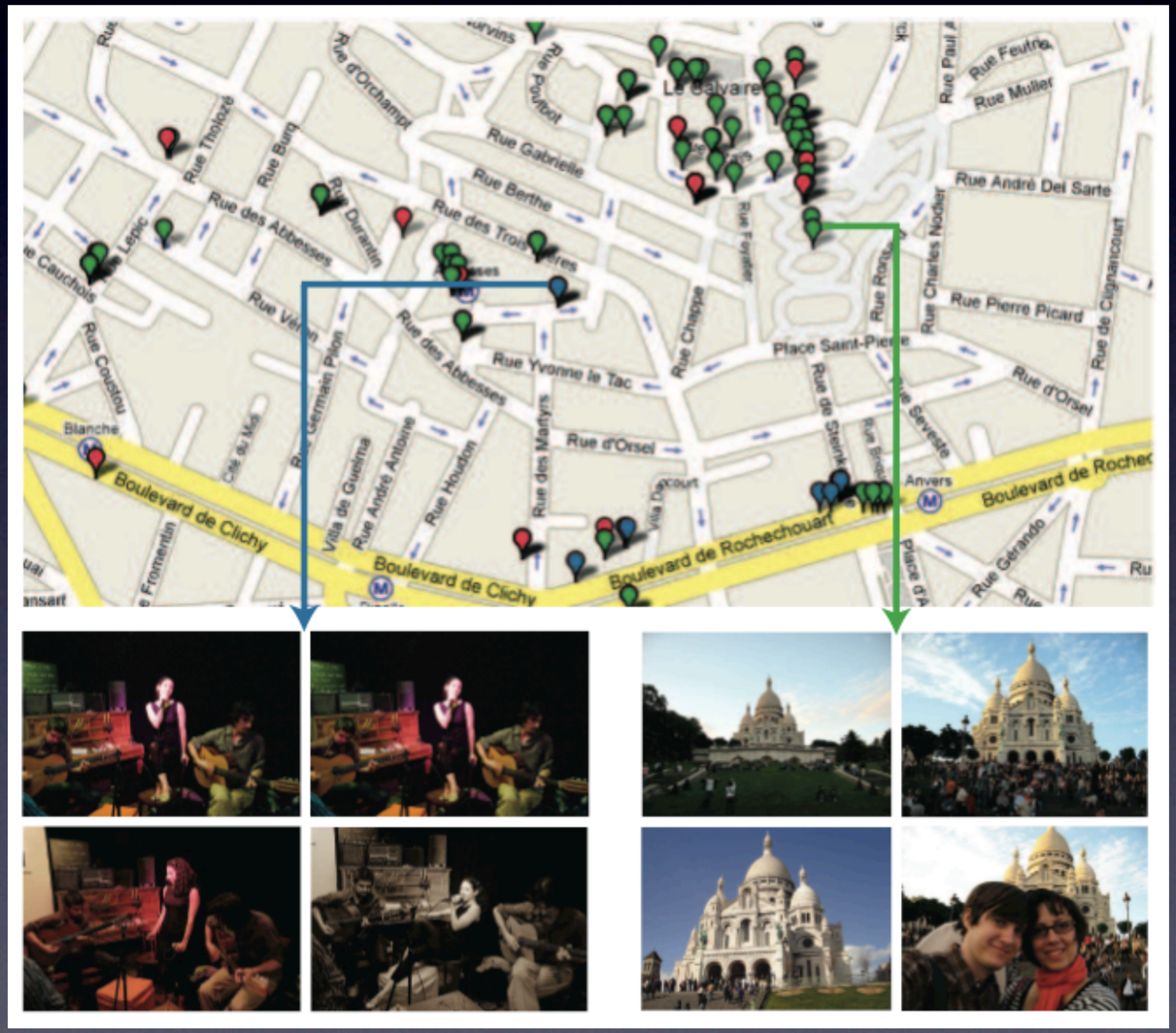
- Precision:
  - objects: **88%**
  - events: **94%**
- Majority of clusters are objects (not astonishing)



- Groundtruth: 700 clusters
  - labeled manually (“object”, “event”, “nothing”)
  - Split: 50% training, 50% test. 10 random partitionings.

Event Cluster Examples



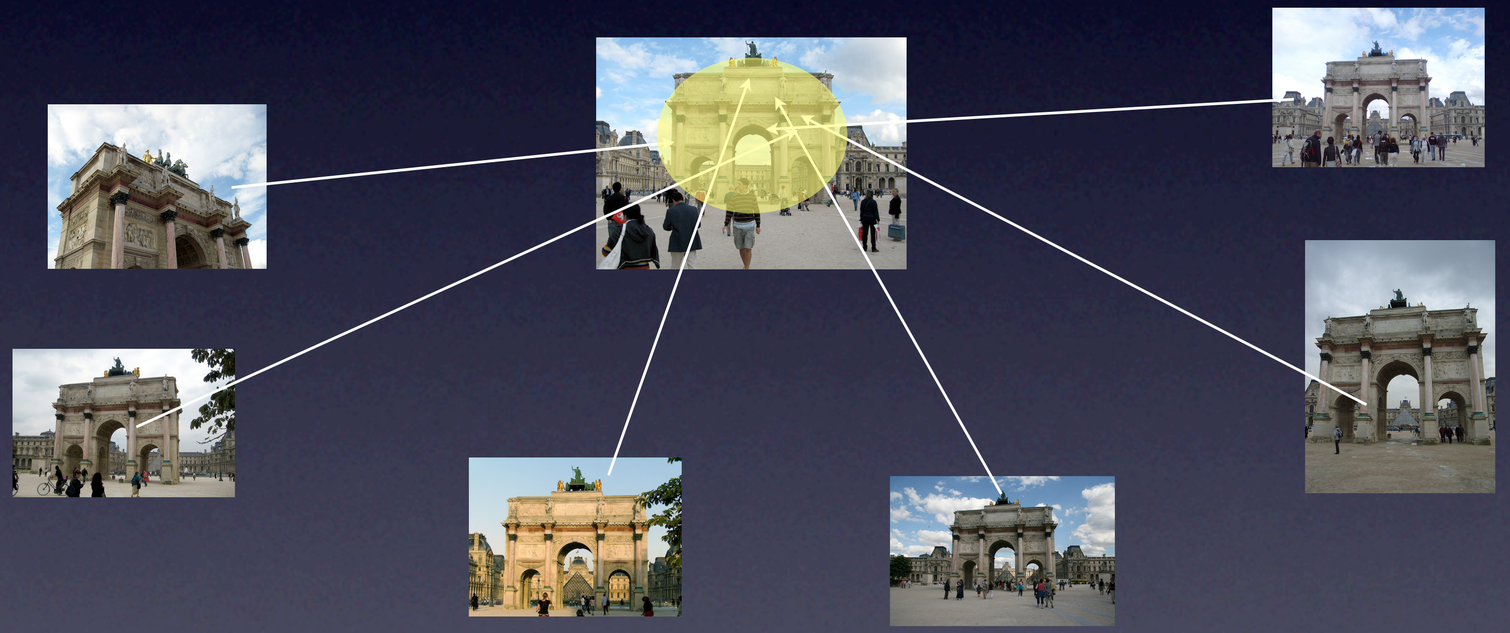


Mining Pipeline Extensions

Extension 1: Object-level mining

Having clusters of many photos for a particular object, allows for "segmenting" the object. *

Enables mining down to the object level, not only image level.



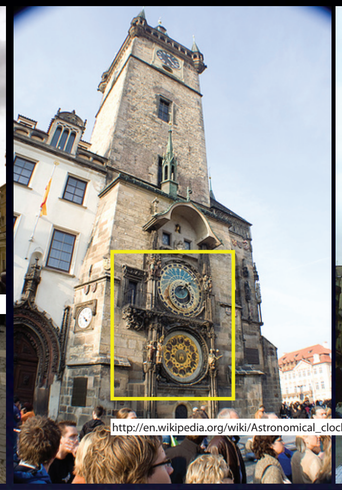
* if same object is photographed from variety of viewpoints, background can be eliminated.

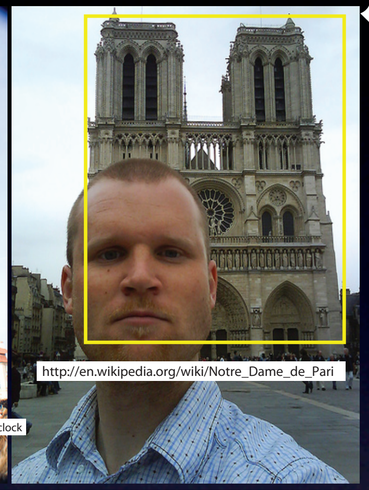
Object-specific feature confidence score

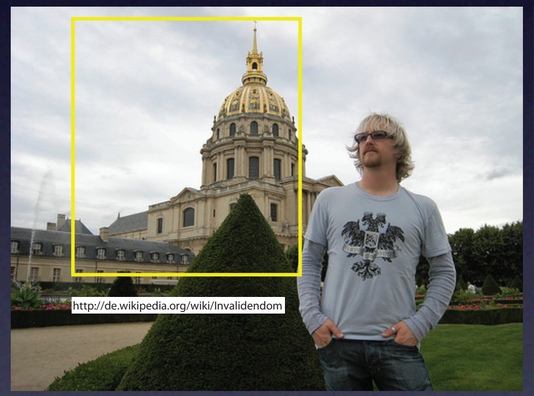


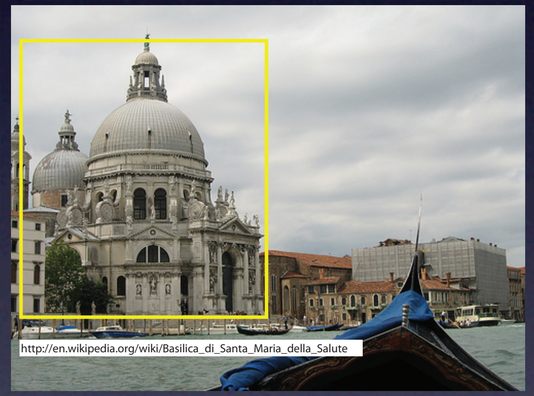
Mined descriptions at object level















Extension 2: growing small clusters

- Recognition or 3d reconstruction more reliable, if
  - more viewpoints of same object in DB
  - more variations (lighting conditions etc.) in DB



- Conflict with crawling based on geo-tagging, which is not general practice (yet)
- Idea: "geo-clusters" as seed, growing with cross-media retrieval?

Cluster growing

Clustering

VISUAL Features



museum, louvre, carrousel, triomphe , arc du triomphe , img_7312, paris...

Frequent Itemset Mining on associated tags (Agrawal 93)

TEXT

museum
museum louvre
carrousel
carrousel triomphe
carrousel triomphe arc
carrousel triomphe arc du
carrousel triomphe du
carrousel arc
carrousel arc du
carrousel du
triomphe
triomphe arc
triomphe arc du
triomphe du
arc
arc du

Google

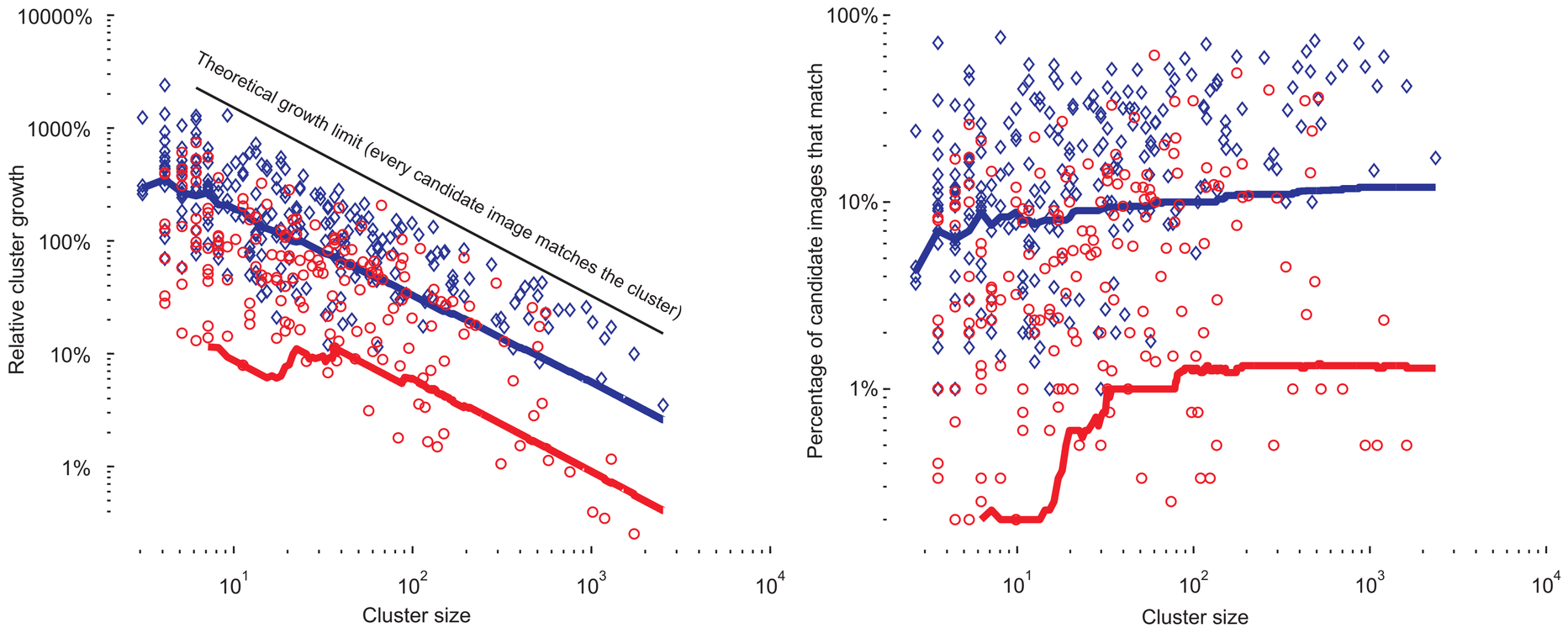
TEXT Retrieval for IMAGES



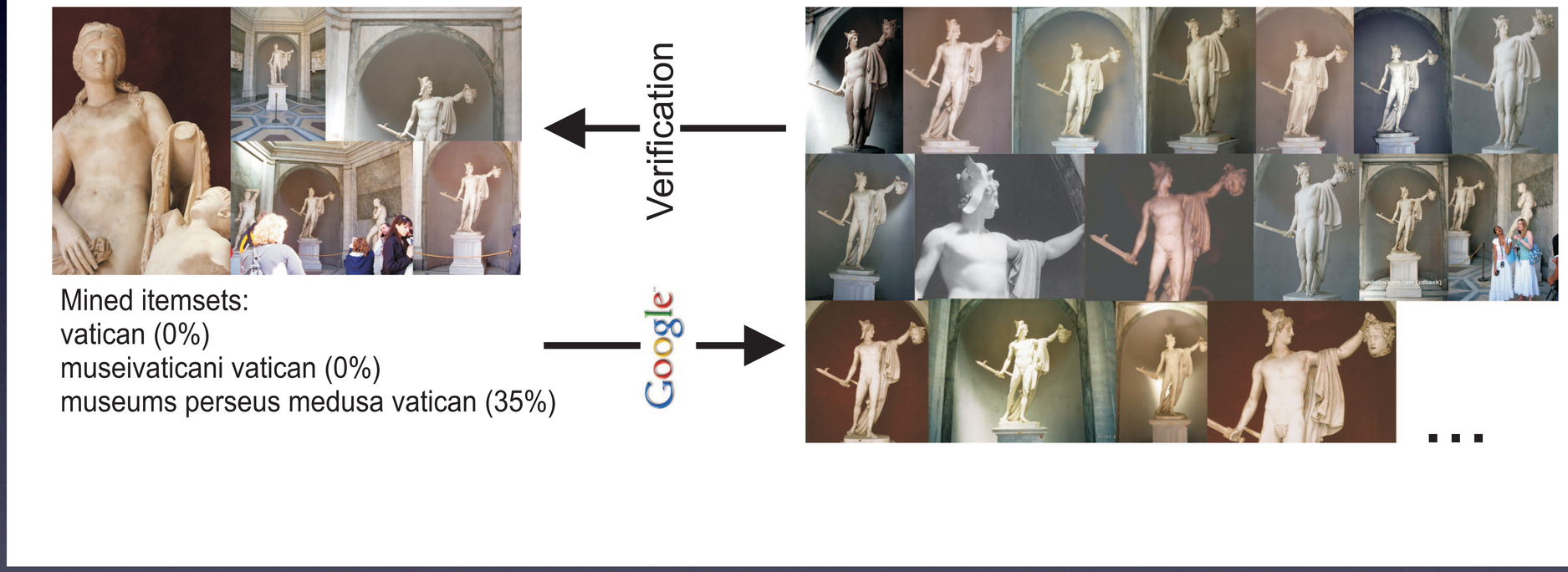
MATCHING by VISUAL Features



Cluster growing stats



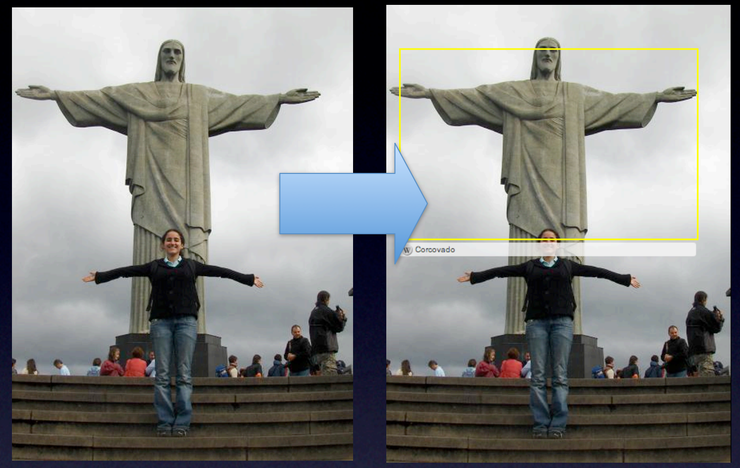
Cluster Growing Example



images in cluster: 5 -> 40.
35 non geo-tagged images added.
Precision *increased* from 80% to 95% due to false positive in original cluster

Applications

Auto-annotation of your holiday snaps



Query Image

Related Content
(e.g. Wikipedia, tags)

Auto-annotating holiday snaps
ICCV September 2009
[Gammeter, Bossard, Quack, Van Gool]

Powerful visual features
&
Scalable Indexing

SURF
[Bay et al. ECCV 06]
Approximate K-Means
[Philbin CVPR 07]

Cross-media reference
database of objects etc.

This Talk
[Quack, Leibe, Van Gool. CIVR 08]

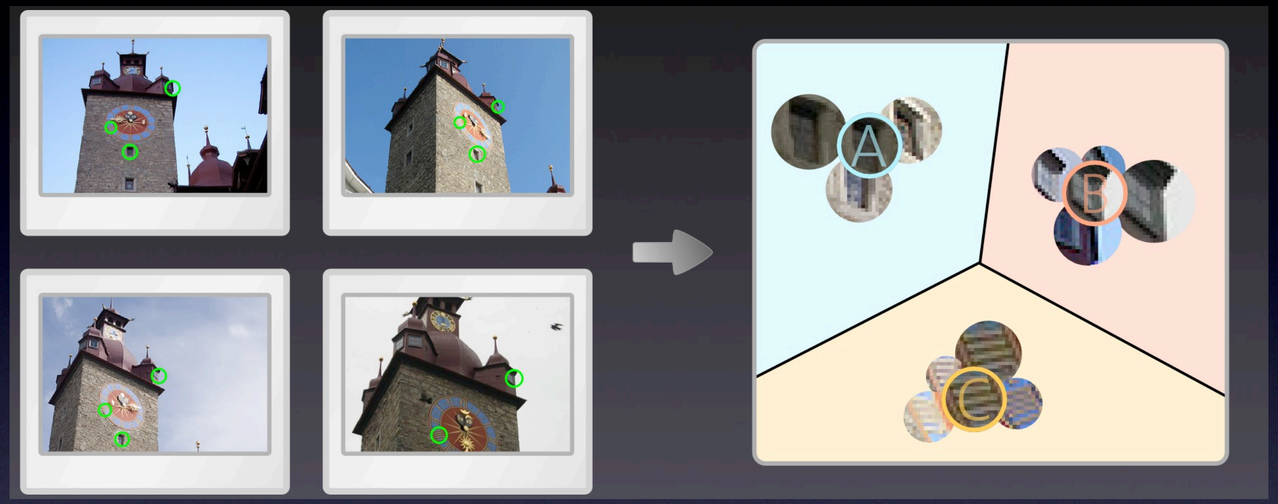
Visual Vocabularies Refresher

Local Features

Appearance

Detectors + Descriptors

[Lowe 01], [Tuytelaars99], [Matas02], [Mikolajczyk04], [Bay06]



Visual Vocabularies

Clustering of feature descriptors into visual words

Feature's appearance summarized by visual word id

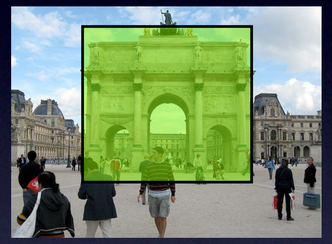
[Burl98],[Weber00],[Agarwal02],[Fergus03], [Sivic03], [Leibe03], [Nister06], [Philbin07]

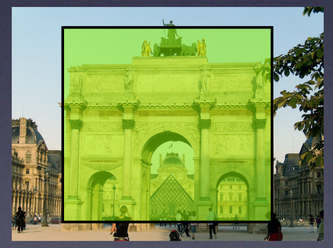


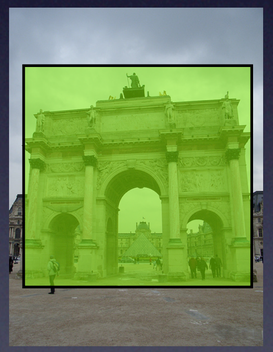
Each image can now be encoded as histogram of visual words.
Standard text retrieval techniques can be applied (inverted index etc.)

Optimized Visual Vocabularies

Using estimated bounding boxes on mined database images we can reduce feature set for indexing



Keep only features from within bounding box.

How much smaller is our index? Do we even gain precision?

Dataset



Crawling “hotspots”

# Images crawled	4’482’582
# Images processed	996’341
# Object Clusters	63’232



Figure 6. Examples from our test dataset.

674 test images

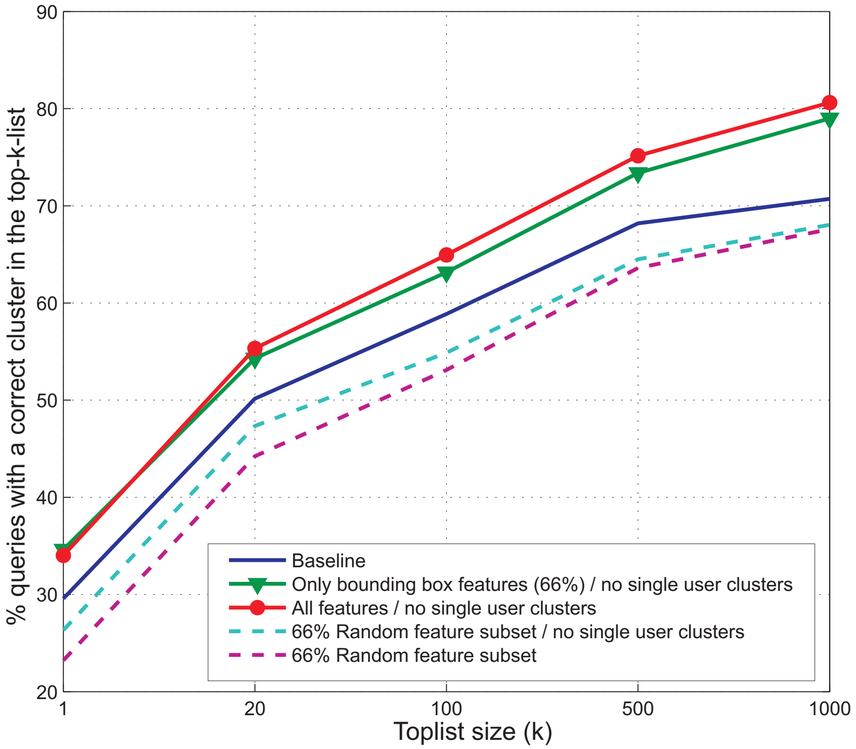
Annotation/Retrieval Results

	Feat. Based	Vis. Word Based
Full Index	452 (67.1%)	324 (48.1%)
BB Features	447 (66.3%)	325 (48.2%)

Final annotation results
(Recognition rate on first rank)

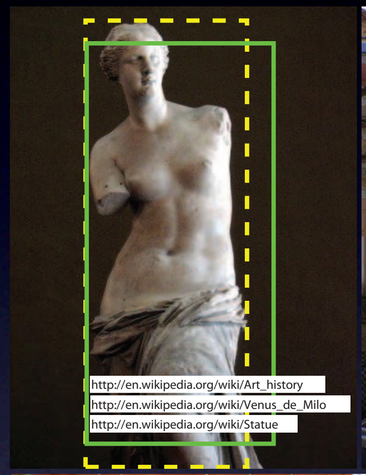
	Uncompressed	Compressed
Full Index	3.7 GB	1.8 GB
BB Features	2.3 GB	1.2GB

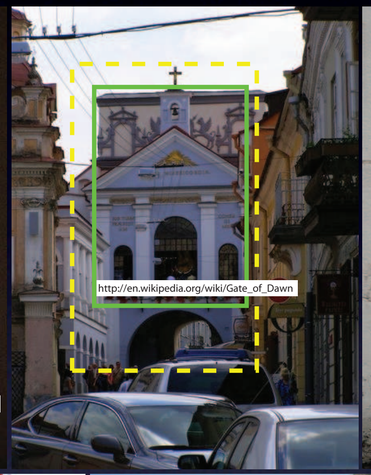
Index sizes
Compression: simple-9



Recognition rate optimized visual vocabularies vs standard visual vocabularies

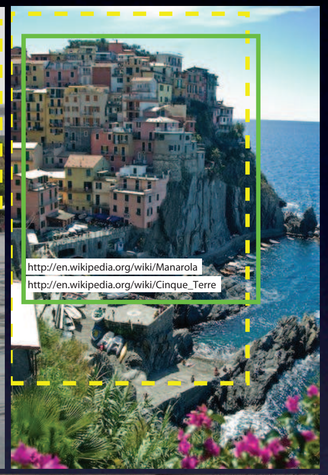
Annotation Results





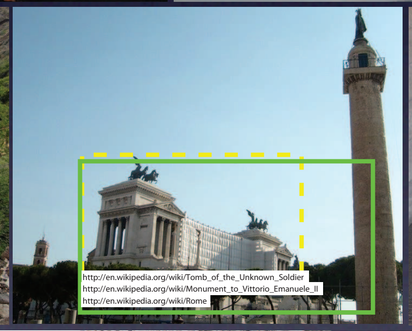


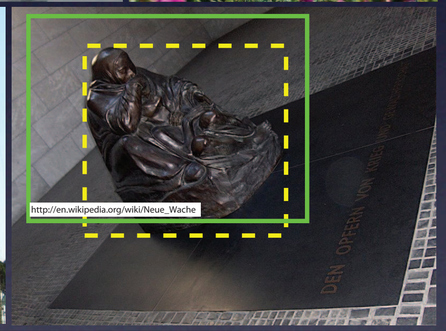




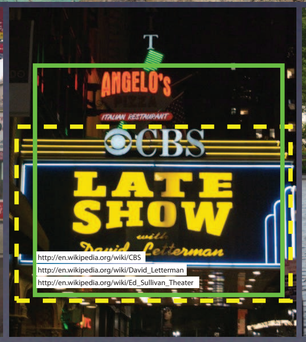








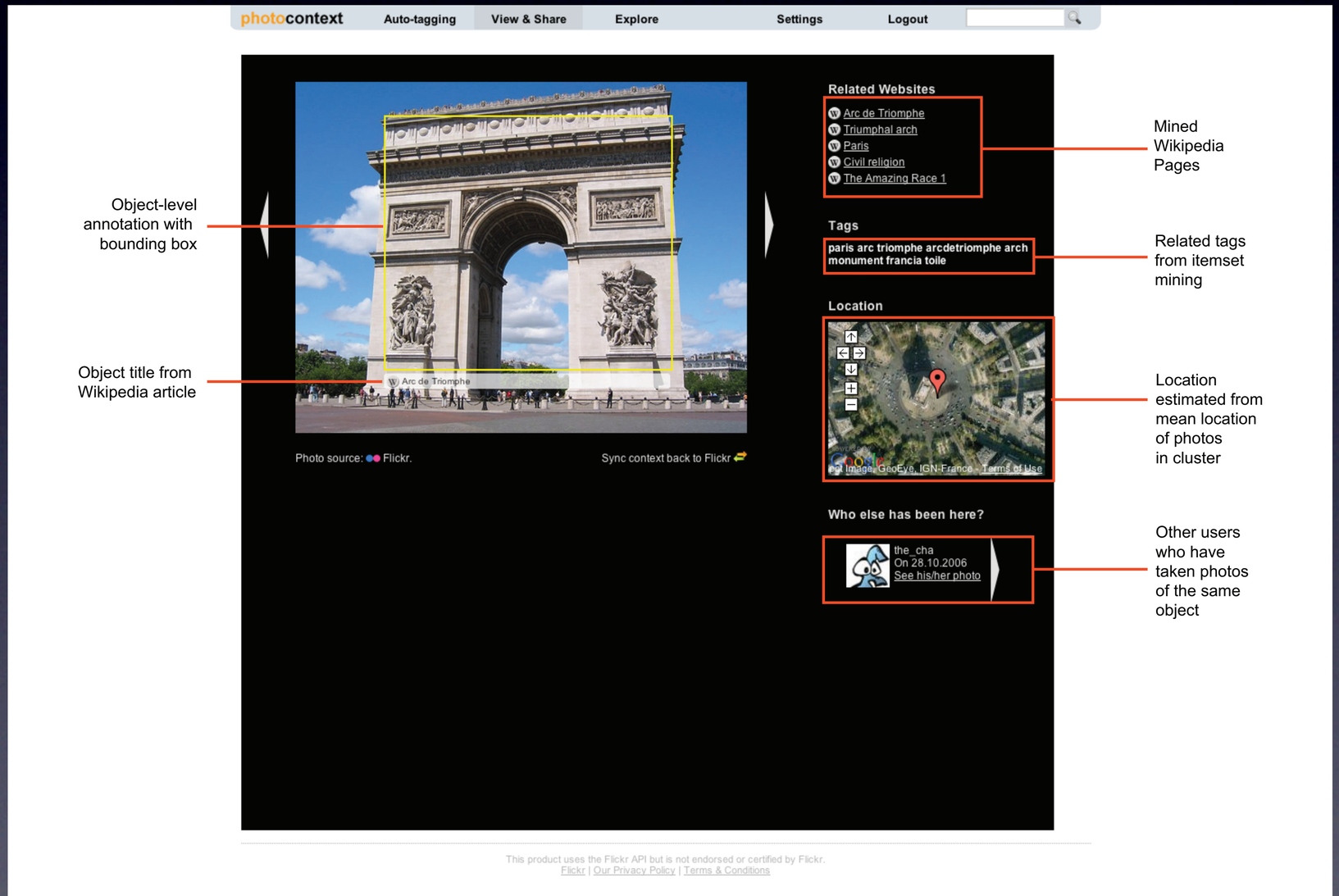


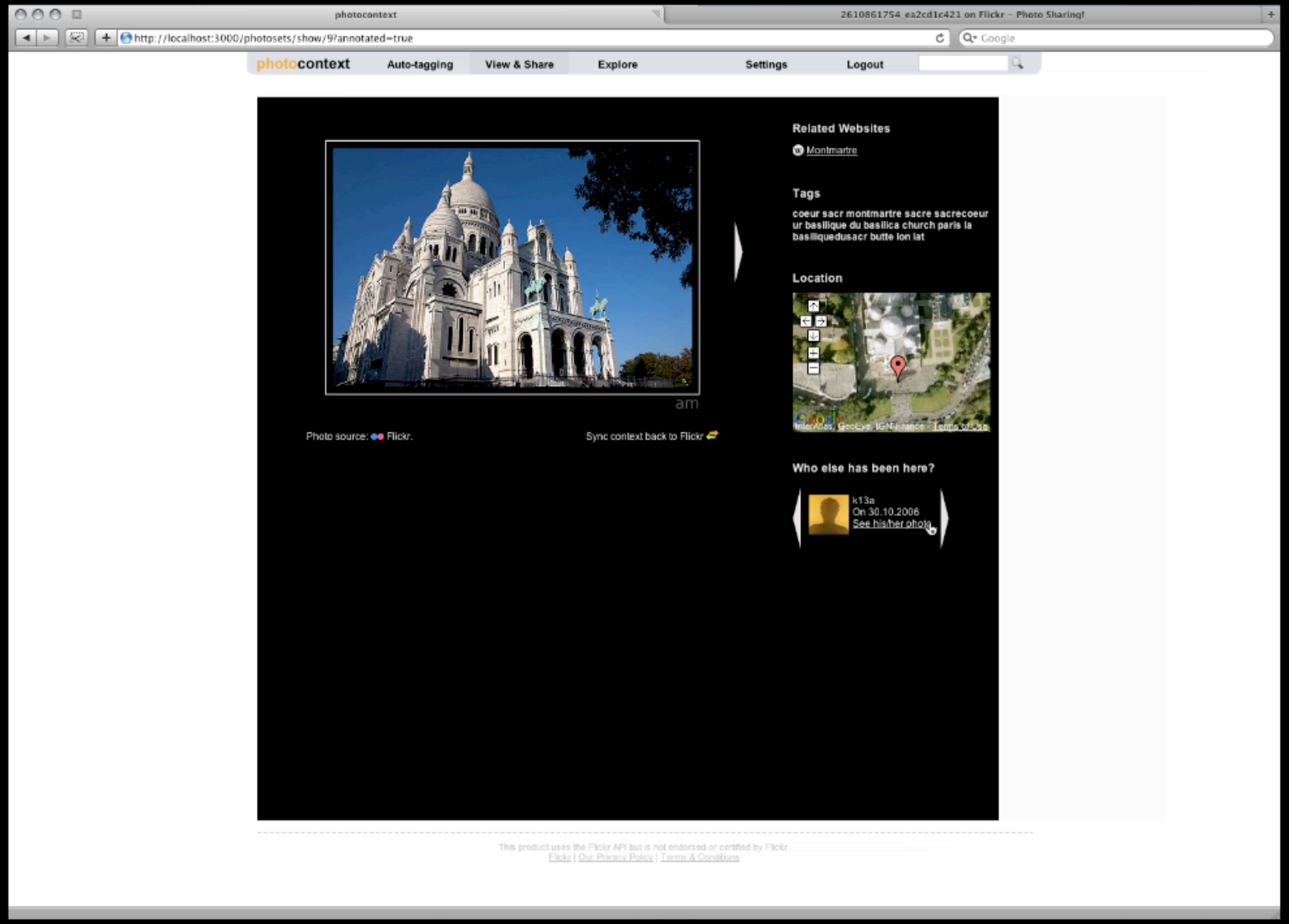






Auto-annotation app prototype





www.autota.gr register for trial access

Augmented Reality



Note: this *really* requires a reference database with object level information

AR Example II



Combination with media cover recognition by kooaba (5 million items) (www.kooaba.com)

Combination with OCR + translation

Conclusions

- Cross-media mining to create reference database for visual retrieval
- Mining pipeline using
  - Geography
  - Visual Features
  - Text
  - Time & User data
- Large-scale experiments (passing 5 million photos currently)
- Applications: prototypes for auto-annotation and augmented reality
  - by combining with large-scale object retrieval system

Future Work

- Database crawling
  - Process more data
  - Better event cluster analysis
- Auto-annotation
  - Faster geometry
  - Model-based indexing
- Applications
  - Public versions

Thanks for listening

auto-annotation demo (online soon):

www.autota.gr

[1] Stephan Gammeter, Lukas Bossard, Till Quack and Luc Van Gool
I know what you did last summer: object-level auto-annotation of holiday snaps (to appear)
ICCV 2009, Kyoto, Japan, 27.9.2009 - 4.10.2009

[2] Luc Van Gool, Michael Breitenstein, Stephan Gammeter, Helmut Grabner, Till Quack
Mining from large image sets
CIVR 2009, Santorini, Greece, 8.7.2009 - 11.7.2009

[3] Till Quack, Bastian Leibe and Luc Van Gool
World-scale Mining of Objects and Events from Community Photo Collections
CIVR 2008, Niagara Falls, Canada, 7.-9. July 2008

tquack@vision.ee.ethz.ch

www.vision.ee.ethz.ch/tquack